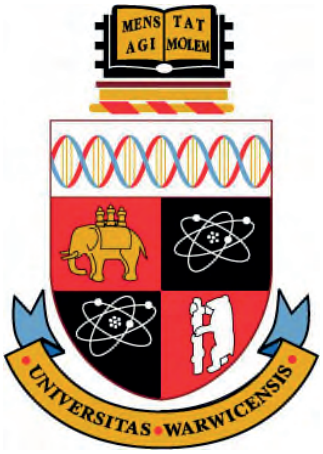


Informative sequence-based models for fragment distributions in ChIP-seq, RNA-seq and ChIP-chip data

Author: Nigel P. Dyer

A thesis submitted to the University of Warwick for the degree of

Doctor of Philosophy

Supervisors: Dr Sascha Ott and Professor Jim Beynon

Molecular Organisation and Assembly in Cells (MOAC)

Doctoral Training Centre

September 2011

Table of contents

† indicates that the section is substantially identical to a section of the paper to be submitted to NAR.

‡ indicates that the section or appendix is substantially identical to a section in the supplementary data of the paper to be submitted to Nucleic Acids Research (NAR).

List of tables

List of figures

Acknowledgements

I would like to thank the following people:

- My wife *Jenny*, without whose support and encouragement this PhD would not have been possible;
- My supervisors, *Dr Sascha Ott* and *Professor Jim Beynon*;
- My children *Naomi* and *Jacob* who encouraged me by regarding it as cool that their dad was doing a PhD in genetics;
- The members of my PhD advisory board: *Professor Richard Napier*, *Dr Andrew Mead* and *Dr Isabelle Carre*; for their interest in what I have been doing, and their support and advice;
- My PhD examiners for their care and attention to detail in reviewing the thesis and for their extensive and extremely helpful comments;
- *Emma Cooke* and *Dr. Katherine Denby* for kindly providing the Arabidopsis RNA-seq data that are analysed in this thesis;
- *Dr. Sally Adams* for kindly providing the Arabidopsis LHY ChIP-seq data that are analysed in this thesis;
- The members of Sascha Ott's group as they have come and gone over the three years of the PhD, with particular thanks to *Boris Noyvert* and *Laura Baxter*, for their many helpful comments and suggestions;
- Other members of the Systems Biology group at Warwick for their helpful support during this work; in particular thanks to *Jay Moore* and *Siren Veflingstad*;
- All of the staff at MOAC for their support and encouragement;
- Professor *Alison Rodger* for her support, encouragement, advice and friendship during my time as a student in MOAC;
- The creators of the cisGenome tool which has proved to be an invaluable foundation for much of the work described in this thesis;
- The Engineering and Physical Sciences Research Council (EPSRC) for providing funding for this thesis.

Finally I would like to thank *John Cook*, who was my boss when I started work at British Telecom Research Labs 31 years ago, and *Roger Mead*, the father of one of my PhD advisors. Their inter-connection clearly requires some explanation. My first job at the Research Labs was to characterise telephone cable networks, a job that required model fitting, and John Cook gave me a model fitting algorithm to use. Since then, my career in research and development has presented me with a series of complex non linear, multi-parameter optimisation problems, including the analogue and digital circuit design in the System X telephone exchange line card [72]. Each time that this has happened I have used the same algorithm. Its most recent application is at the heart of this PhD and it was only recently, with John Cook's help many years after he first suggested that I use the algorithm, that I found that it was the Nelder-Mead algorithm [74]. Roger Mead, one of the co-authors, is the father of Andrew Mead who sits on my PhD advisory panel.

Declaration

This thesis is presented in accordance with the regulations for the degree of Doctor of Philosophy. No part of this work has been previously submitted for another degree. At the time of submission of this thesis, some sections of this work were in the process of being submitted for publication. These sections are indicated by the symbols † and ‡ in the headings. Appendix G provides details of papers that were published during the preparation of the PhD for which I was a co-author. These are peripheral to the main thesis. My contribution to these papers and the relationship of this work to the thesis are identified in Appendix G.

Nigel Dyer
September 2011

Abstract

Many high throughput sequencing protocols for RNA and DNA require that the polynucleic acid is fragmented so that the identity of a limited number of nucleic acids of one or both of the ends of the fragments can be determined by sequencing. The nucleic acid sequence allows the fragment to be located within the genome, and the fragment distribution can then be used for a variety of different purposes. In the case of DNA this includes identifying the locations where specific proteins are bound to the genome. In the case of RNA this includes quantifying the expression levels of different gene variants or transcripts. If the locations of the polynucleic acid fragments are partly determined by the underlying nucleic acid sequence this could bias any results derived from the data. Unfortunately, such sequence dependencies have already been observed in the distribution of both RNA and DNA fragments. Previous analyses of such data in order to reduce the bias have examined the role of regional characteristics such as GC bias, or the bias towards a specific sequence at the start of the fragments.

This thesis introduces a new method for modelling the bias which considers the degree to which the nucleotide sequence affects the likelihood of a fragment originating at that location. This shows that there is often not a single bias characteristic, but multiple, alternative sequence biases that coexist within a single dataset. This also shows that the nucleotide sequence immediately proximal to the fragment also has a significant effect on the fragment likelihood. This new approach highlights characteristics that were previously hidden and provides a more powerful basis for correcting such bias.

Multiple alternative sequence biases are observed when both RNA and DNA are fragmented, but the more detailed information provided by the new technique shows in detail how the characteristics are different for RNA and DNA and indicates that very different molecular mechanisms are responsible for the biases in the two processes.

This thesis also shows how removing the effect of this bias in ChIP-seq experiments can reveal more subtle features of the distribution of the fragments. This can provide information on the nature of the binding between proteins and the DNA with per-nucleotide precision, revealed through the change in likelihood of the DNA fragmenting at each position in the binding site.

It is also shown how the model fitting technique developed to analyse sequence bias can also be used to obtain additional information from the results of ChIP-chip experiments. The approach is used to find the nucleotide sequence preference of DNA binding proteins, and also the cooperative effects associated with binding at multiple binding sites in close proximity.

Abbreviations

3D	Three Dimensional
DNA	Deoxyribonucleic Acid
CD4	Cluster of Differentiation 4: A glycoprotein expressed on the surface of certain cells
cDNA	complementary DNA: DNA that has been synthesized from messenger RNA
CPU	Central Processing Unit
EM	Expectation Maximization
ENCODE	Encyclopaedia of DNA Elements
EPSRC	Engineering and Physical Sciences Research Council
GABP	Growth-associated Binding Protein
GEO	Gene Expression Omnibus
GUI	Graphical User Interface
IP	Immunoprecipitated
MEME	Multiple Expectation Maximization for Motif Elicitation
NCBI	National Centre for Biotechnology Information
mRNA	Messenger RNA
NAR	Nucleic Acids Research
NRSF	Neuron-restrictive Silencer Factor
oriC	Origin of chromosomal replication
PCAF	P300/CBP-Associated Factor: A transcription factor
PCR	Polymerase Chain Reaction
PSSM	Position Specific Scoring Matrix
PCM	Position Coefficient Matrix
QuEST	Quantitative Enrichment of Sequence Tags
RNA	Ribonucleic Acid
SRA	Sequence Read Archive
SNP	Single Nucleotide Polymorphisms

Glossary

ChIP-chip	Chromatin Immunoprecipitation followed by DNA quantification using microarray technology (chip) which is used to determine protein binding to DNA
ChIP-seq	Chromatin Immunoprecipitation followed by sequencing, which is used to determine protein binding to DNA.
CD4 cells	A type of lymphocyte (white blood cell) cell with a CD4 receptor,
HeLa	A common immortal human cell line derived from a cancer cell line taken from Henrietta Lacks.
K562	An immortalised human cell line created from myelogenous leukaemia cells.
TRANSFAC	A database of eukaryotic transcription factors, their experimentally proven binding sites, and regulated genes.

Chapter 1

Introduction

1.1 Motivation and overview

A way of introducing the motivation for the work described in this thesis is by reference to the graphs shown in Figure 1-1. These show the results from a ChIP-seq experiment whose purpose was to identify the locations in the genome where specific proteins were bound. During ChIP-seq experiments DNA is extracted from cells, fragmented, and the ends of the fragments are sequenced, generating the sequence 'tags' which are used to identify from where in the genome they came.

Figure 1-1 is a histogram of the data for a short region of the genome, with each bar indicating the number of fragments that were found to start (upper graph) or finish (lower graph) with respect to the forward strand direction at each genomic coordinate. In such experiments the data are then used to identify the regions where the target protein was originally bound. Various techniques and algorithms have been developed to identify these regions using the data but at the heart of all of the algorithms some form of averaging is applied to the data to remove what appears to be noise in the data and so create a smoothed or averaged version of the distributions. The various algorithms then use the smoothed data to determine the likely locations where the proteins were bound.

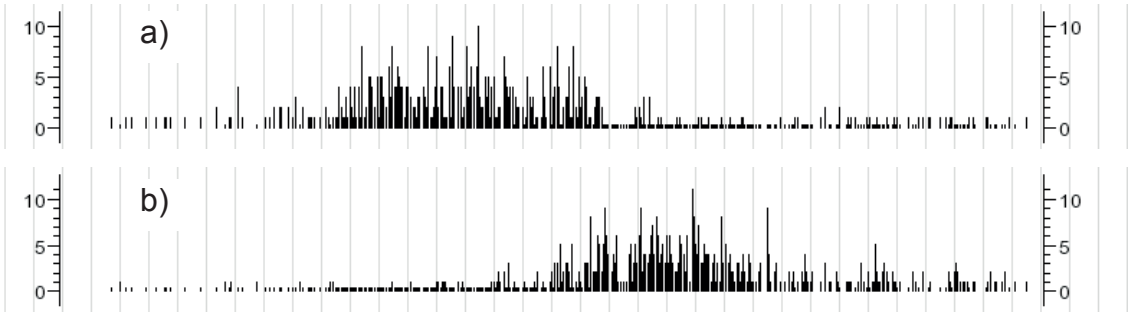


Figure 1-1 Example ChIP-seq data. a) The distribution of fragment starts with respect to the direction of the forward strand in a region of the genome. b) Distribution of fragment ends. Each bar indicates the number of fragments associated with a single genomic coordinate.

The motivation for the research described in this thesis was a growing suspicion that by averaging the data in this way, information was being discarded that could otherwise have been used to provide more details about the binding of proteins to the DNA. The initial

suspicion that this was the case arose because of apparent similarities in the data from the same region in similar experiments. This implied a degree of correlation in the relationship between fragment start position and local sequence and suggested that the sequence tags contained information at the level of individual nucleotides, and therefore the seemingly random variation in the fragment counts in Figure 1-1 is not simply noise.

This led to a number of interrelated avenues of research, each of which is covered in separate chapters within this thesis. These avenues confirmed that there is additional information that can be obtained from the fragment distribution and that this can shed light on a number of aspects of both ChIP-seq and RNA-seq experiments. In order to analyse these effects, tools were developed based on model fitting. This approach was then found to have the potential for wider application in interpreting the results of ChIP-chip experiments. The tools and some results from analysing ChIP-chip experiments are also described.

1.2 Overview of the document structure

The core of this thesis consists of six chapters, starting with this first introductory chapter which provides the background to the work described in the thesis and proceeding through to a concluding chapter which draws the various threads together. Between the two are a series of four relatively self contained chapters describing various aspects of the work that has been carried out, each of which has an 'introduction, results, discussion' format that deals with the specific topic. In some cases there is a short discussion section directly associated with specific results. This is done when the discussion is largely unrelated to the rest of the chapter, or provides key ideas that form the basis of work that is described in the following results sections.

1.2.1 Relationship to published papers

The information presented in Chapters 2 and 3 has also been covered in an article together with extensive supplementary data that has been submitted to Nucleic Acids Research (NAR). Those sections of the thesis that align closely to text in the paper itself are marked †, and those sections of the text that align with text in the supplementary data for this paper are marked ‡.

1.2.2 Appendices

Appendix A describes the development of an algorithm and associated code to locate regions in the genome where the sequence matches a sequence elsewhere in the genome. This

information was required in order to carry out the analysis in Chapter 2, although the details of how this was carried out are not relevant to the main body of the thesis. This work was carried out because at the time this information was required there was no generally available tool for doing this.

Appendix B and C are supplementary material for Chapter 2, providing more analysis of ChIP-seq data to support the analysis within this chapter. Appendix D is supplementary material for Chapter 3 and provides an analysis of more RNA-seq data to support this chapter.

Appendix E provides a general introduction to the software methodologies and architecture that were used during the work described in this thesis. Appendix F is referenced at various points in the thesis and provides supporting information for some of the algorithms and mathematical analysis that were used. Finally, Appendix G provides a brief summary of the contributions by the author to various journal publications that were prepared during the research described in this thesis, and describes the relationship between the contribution to the papers and the thesis.

1.3 Background

For many centuries, one of the core mysteries in the study of living organisms was the mechanism by which the information that determined the nature of an organism was passed from one generation to the next, and how the information was then used to control the growth and functioning of each successive generation of organisms. The foundation for the solution to this mystery was laid in the 1660s when Robert Hooke was first able to show that living organisms were composed of very large numbers of minute cells, sharing many features in common but nevertheless capable of being tailored to carry out the many different roles required in even the most basic of multicellular organisms [41].

While this laid the foundation, it also demonstrated the challenge faced by anyone attempting to solve this mystery, in that Hooke's observations and later interpretations of these observations showed that the information that was being sought was wholly contained within these microscopic cells.

It was perhaps the pioneering work of Theodor Boveri over 200 years later that provided the next significant contribution to solving this puzzle. He was able to show that it was the chromosomes within the cell nucleus which were responsible for carrying the information from one generation of organism to the next, and from one generation of cell to the next within an organism as the cells underwent cell division [14]. He was also able to

show that the information relating to individual characteristics was consistently contained within specific chromosomes.

At about the same time, the rules of inheritance were being determined through observations of the variation in the characteristics of an individual from one generation to the next within a species. This lead to the idea of well defined units of inherited information, for which the term gene was introduced in 1909 by Wilhelm Johannsen [48].

The solution to the problem of how the abstract genetic information that is associated with the gene was connected to the physical structure of the cell came in 1941 with the concept of "one-gene-one-enzyme" where it was proposed that each of the different enzymes in the cell was associated with a one unit of genetic inheritance [10]. It has been observed that this finally began to connect genetics, which had previously been a specialised science with its own language and ideas, to the rest of the biological sciences [43].

The pivotal insight that showed how the genetic information could be stored and replicated within the chromosomes was the solving of the structure of DNA by Crick and Watson [100, 101]. Subsequently, Crick was able to develop this insight into what he termed the central dogma of molecular biology, the transfer of the genetic information in the DNA to RNA which then determines the sequence of the proteins. This concept is still at the heart of molecular biology [22].

The details of this dogma have continued to be worked out, including the identification of the locations of the genes within the genome to an ever greater precision. This led to the realisation that in eukaryotes, the information that is used to determine the amino acid sequence of the protein is not held in one continuous sequence, but is encoded in a series of short sections or exons, with extended regions or introns in between [11, 12].

The intron-exon structure of genes highlighted one of a growing number of areas where the simple 'one-gene-one-enzyme' concept is a simplistic picture of a significantly more complex reality. For example, introns and exons provide the flexibility of being able to select different combinations of exons to create different RNA transcripts in order to produce variants of the protein which can be tailored to different situations where the protein is required [Reviewed in 75].

There remained many questions, including the question of how many genes there were in the genome, the answer to which would help identify how much information is available within the genes to determine the complexity of the structure and function of multicellular organisms.

The advent of high throughput sequencing, which enabled almost the complete DNA sequence of organisms with complex genomes such as *Homo sapiens* to be determined [56], started to provide answers to this question and significantly, the answer of about 25000 genes was much lower than expected [1]. Furthermore, the significant similarity between many of the genes in different organisms was increasingly showing that the protein coding sequences themselves only contained a small part of the genomic information that determines the structure and function of an organism.

Genes can be considered as being like the instruments of an orchestra which, with essentially the same set of instruments, is capable of playing a vast repertoire of music, from Bach to Bacharach. The structure and character of the music being determined by the order in which the instruments are played, and how they are played.

The same is true of genes. It was increasingly clear that it is the order and degree to which the genes are expressed that is critical, and this is determined by a complex and often subtle network of different interacting components. One of the first aspects of this control network to be examined was the binding of proteins, commonly known as transcription factors, to the region immediately adjacent to the start of the gene where they had a major role in determining when an how much a gene was transcribed into RNA [For a review see 27]. The locations where such proteins bind are commonly referred to as transcription factor binding sites (TFBS). As well as proteins binding to DNA, it became clear that there were many other mechanisms that had a significant role in controlling gene expression. For example, the modification of the DNA itself, often through methylation, also had a role in controlling gene expression, and this was often intimately linked to variation in the structure of the DNA, which is also very important [42, 102]. It was also clear that the RNA, as well as being the intermediate in the transfer of information from DNA to protein sequence, also has a very significant regulatory role with the discovery in 1993 of microRNAs. These are small lengths of RNA that are between 20 and 25 nucleotides in length which are transcribed from many regions of the genome and which then able to control various aspects of the transcription and translation process [59, 80].

At the same time as it was becoming clear that the process of gene regulation was incredibly complex, the continuing development of high throughput sequencers now provided significantly more data from a range of different processes which incorporated sequencing which could be used to help unravel this complexity [53].

Although the binding of proteins to DNA was one of the earliest methods of gene regulation to be investigated, the details of the gene regulation networks associated with such

protein binding are still poorly understood and high throughput sequencing continues to be widely used to investigate protein binding to DNA. The classic high throughput sequencing process that is used for this is ChIP-seq (Chromatin Immunoprecipitation followed by sequencing) [77]. Another approach, based on microarray technology is ChIP-chip (Chromatin Immunoprecipitation followed by microarray (chip)) procedure. This thesis is concerned with extending the techniques for interpreting the data associated with both of these procedures, so they will both be described in more detail in the following sections.

Next generation sequencing is also widely used to investigate the different RNA transcript variants that are generated from the same gene, using the RNA-seq protocol. This thesis is also concerned with how the extension of the techniques used to interpret ChIP-seq data can also be applied to RNA-seq data, so this process will be introduced in more detail in Section 1.5.

1.4 An introduction to the ChIP-chip and ChIP-seq protocols

Chapter 2 and Chapter 4 both relate in some way to the ChIP-seq procedure [77] whereas chapter 5 relates to the use of the ChIP-chip procedure.

The power and range of these processes have resulted in a vast literature describing these techniques and the results that have been obtained using these techniques. The following is a brief introduction to them, concentrating on some of the aspects of these techniques that are particularly important when considering the results in later sections of this thesis.

1.4.1 The motivation for studying protein binding to DNA

The primary purposes of the ChIP-seq and ChIP-chip protocols are to locate the regions in the genome where specific proteins are bound and to quantify the degree of binding at these locations. There are many reasons for wanting to identify where the proteins are bound. For example, it is frequently the case that such proteins are regulating the transcription of genes, in which case they are known as transcription factors. Knowledge of the binding sites can therefore provide information about the identity of the genes that are regulated by a protein [Reviewed in 7]. Once binding sites have been located these can then be used to determine the DNA sequences to which the protein tends to bind [For reviews see 4, 23, 54]. With such knowledge it is then possible to predict binding sites '*in silico*' rather than having to determine them experimentally, for example when investigating variation in gene regulation in a set of organisms which share a common transcription factor [9].

As well as using this information to try and understand gene regulatory networks in different species these techniques have also been used to ascertain the extent of the variation between individuals in a species in order to understand how this variation might explain the specific characteristics of the individuals [e.g. 50].

Gene regulation is a very dynamic process, and both ChIP-chip and ChIP-seq are used to measure the variation in time of protein binding to the DNA in order to understand how the changes are linked to changes in response to external stimuli [e.g. 61], or other signalling pathways [e.g 29], or changes over time as a result of clock networks within living organisms [e.g. 19]. The changes in protein binding can also be associated with changes in gene expression as part of the process of identifying the myriad of different signalling pathways within cells.

As well as proteins playing a role in gene regulation, there is a growing awareness of the role of the chromatin structure in gene regulation, and the use of ChIP-seq and ChIP-chip protocols to locate where structural proteins such as histones bind to the DNA can provide significant information on the chromatin structure [65].

1.4.2 Preparing the DNA for ChIP-seq and ChIP-chip: protein fixing with formaldehyde

In both ChIP-seq and ChIP-chip processes the DNA has to be extracted and purified from the cell while the proteins are still bound to the DNA in order that their presence can be used to identify the protein binding sites. In order to ensure that the proteins remain bound, one of the first stages in the protocol is to fix the proteins to the DNA using formaldehyde.

Formaldehyde reacts with both amino groups in proteins and also DNA to form a Schiff base, which indicates that a carbon-nitrogen double bond is formed during the reaction. This bond is the starting point of a subsequent reaction with second amino group which results in a DNA/protein crosslink or a link between amino acids within the same or two different proteins. The degree of cross linking is important for satisfactory completion of later stages of the protocol in that both excessive and inadequate cross linking can result in poor yield [76]. The advantage of using formaldehyde is that the process is believed to be completely reversible, allowing the proteins to be released from the DNA at a later stage so that the DNA can be sequenced [76].

1.4.3 DNA extraction and fragmentation

Following DNA extraction from the cell, the DNA is then fragmented into lengths that are of the order 100 to 300 base-pairs, typically using sonication. This is perhaps the least well controlled of the stages in the process because of the difficulty of applying the same amount of sonication to the sample in successive experiments, and the difficulty in ensuring that all of the sample is equally affected by the sonication.

Sonication is a well established process for disrupting biological material such as cells, and there are a number of types of standard laboratory sonicators that have been used for many years. There are a variety of different approaches to the problem of applying ultrasound to the samples, including placing a tip which is vibrated at ultrasonic frequencies into the sample or placing multiple samples into a water-bath that is then excited with ultrasonic vibrations which are carried through the water to the sample (Figure 1-2).

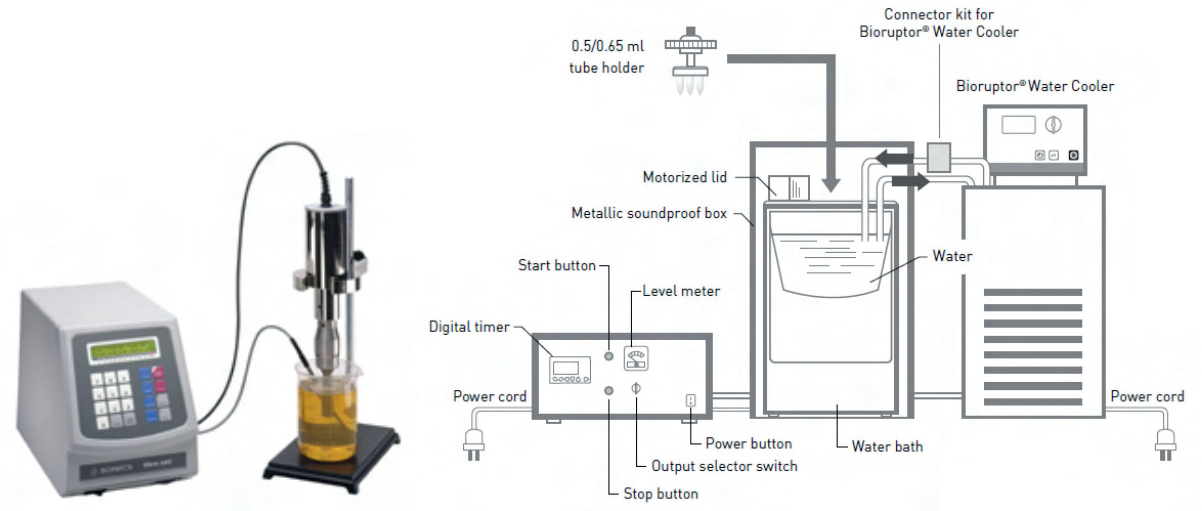


Figure 1-2 Examples of laboratory sonicators. a) Tip based sonicator where a tip is placed into the liquid to be sonicated (www.sonics.com) b) The Bioruptor® (www.diagenode.com). An example of systems where multiple samples are sealed in individual tubes and immersed in a water bath through which the ultrasound is carried to the samples.

Products such as the Bioruptor® from Diagenode have been designed specifically for the task of fragmenting DNA samples, and are designed to try and ensure that the ultrasound is evenly distributed through all of the samples, and that the samples are kept cool.

Ultrasound does not fragment the DNA as a result of direct coupling of the acoustic wave to the DNA as the acoustic wavelength is greater than 1 cm and so unable to couple directly to the DNA whose dimensions are many orders of magnitude smaller. Ultrasound can cause damage to the cell as a result of bulk heating, but DNA fragmentation protocols attempt

to minimise this by using short bursts of ultrasound interspersed with cooling on ice, or by actively cooling the water in water-bath based sonicators.

It is generally understood that DNA fragmentation occurs as a result of cavitation that occurs within the sample [69, 93]. Microscopic bubbles form and grow within the solution during the negative pressure half-cycle of the pressure wave, and in the positive half cycle they reduce in size. If they grow beyond a certain critical size then they collapse catastrophically resulting in a very intense localised release of energy. DNA fragmentation is believed to occur largely as a result of local shear stresses that occur when the bubbles collapse, although the effect of localised heating may also be significant as temperatures can briefly reach 5000 ºC [93].

Bubble collapse also produces free radicals, and it has been suggested that these may also play a role in fragmenting the DNA. It is thought that such free radicals will cleave the DNA at random locations whereas cleavage resulting from shear stresses is likely to reduce the DNA fragment size by a progressive series of halvings [35].

It is known that the cavitation that occurs during sonication is dependent on the dissolved gases in the solution [71] which may explain why it has been observed that the degree of fragmentation that can be achieved can be a function of the dissolved gases that are present in the sample [28].

The sonication typically produces fragments over a wide range of lengths, but many DNA sequencing or microarray technologies perform best with fragments that cover a relatively narrow range of lengths. It is therefore necessary to tune the sonication process so that the peak in the distribution of fragment lengths that are obtained is roughly in the region preferred by the subsequent stages of the process, which is typically of the order of 200 to 400 basepairs.

1.4.4 Immunoprecipitation and size selection

The next stage, which is also common to both protocols, is to separate the DNA fragments to which the target protein are bound from the remainder of the fragments. This is done with an antibody for the target protein which is typically attached to a magnetic bead. The DNA with the attached target protein binds to the antibodies on the beads allowing the other DNA to be washed away. In practise, some DNA where there is no bound protein will also attach to the beads. The bound DNA fragments are then washed off the beads, and the formaldehyde cross linking removed leaving free DNA fragments.

In order that the fragments can ultimately be sequenced, it is necessary to ligate adaptors onto the end of the fragments such that the adaptors at each end of each single stranded fragment have different sequences and are not simply complements of each other. If a double stranded adaptor was ligated onto the ends of the fragments then the adaptors at each end of the single stranded DNA would be complements of each other. In the case of the Illumina protocol, the use of Y adaptors, which are double stranded for part of their length and single stranded for the rest, ensures that, after amplification, the adaptors at each end of the fragment have different sequences.

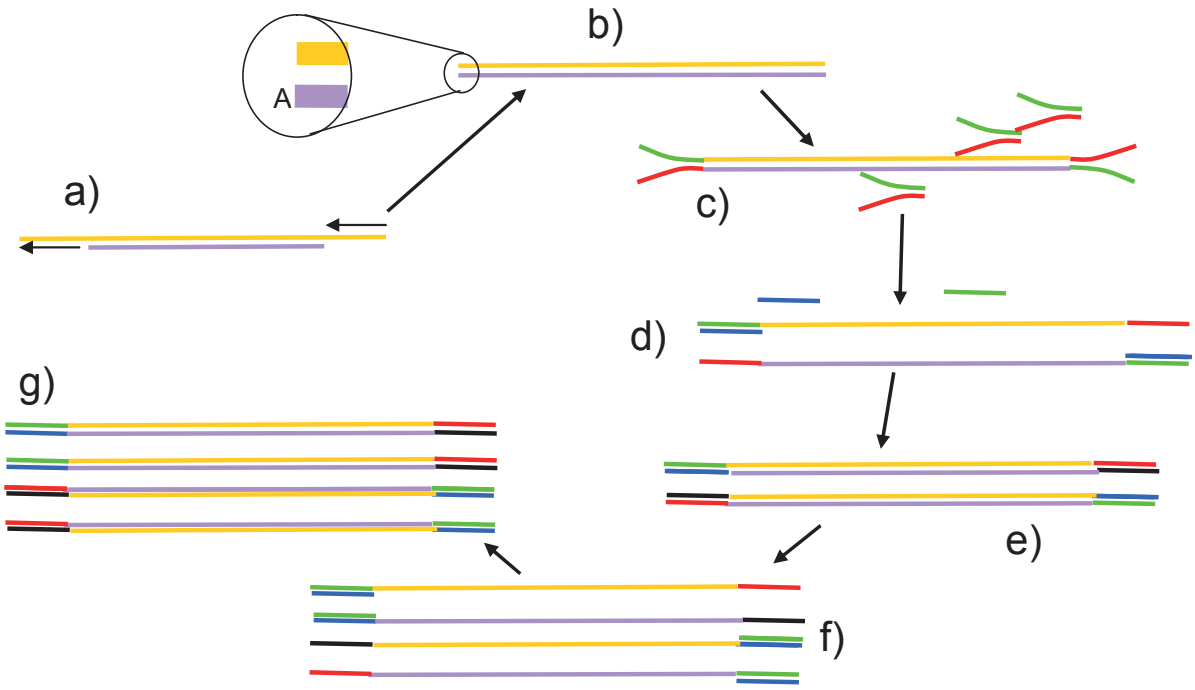


Figure 1-3 Processing and amplification of DNA fragments for use in the Illumina sequencer. a) Double stranded DNA fragments are repaired by extending 3' excessive ends and digesting 3' protruding ends. b) Blunt ended strands to which an A overhang is added to the 3' end of each strand. c) This allows a 'Y' adaptor to be ligated onto both ends which starts with a short section of double stranded DNA and continues with two sections of uncomplementary single stranded DNA. d) These are denatured and then allowed to bind to two added primers (green and blue), although at this stage only one of them is able to bind to the single stranded adaptor sequences. e) The second DNA strands are synthesised. f) The strands are denatured, and at this stage both primers are able to bind to the end adaptor sequences. g) The process continues, creating a large number of identical fragments with different adaptor sequences at each end.

Before the adaptors are ligated onto the fragment ends, the fragments are repaired. The 'Y' adaptors are then ligated onto the ends of the fragments, which are then amplified using polymerase chain reaction (PCR). Figure 1-3 shows how the use of 'Y' adaptors ensures that

the adaptor sequence at each end of the fragments after amplification are different. In the figure a green and red adaptor, with different nucleotide sequences, end up on either end of one of the strands of the DNA, and their complementary red and blue adaptors on the ends of the other strand.

When the fragments are amplified for the ChIP-chip process a conventional fully double stranded adaptor is used as there is no requirement for the adaptors at either end to have different sequences.

As well as amplification the fragments are size selected, where the DNA is run on an agarose gel and a strip of gel is excised that corresponds to the optimal range of DNA lengths. The end of the process yields a sample of fragments at an adequate concentration for quantification using either microarray (ChIP-chip) or sequencing (ChIP-chip) technologies.

1.4.5 The use of input DNA or mock precipitated DNA as a control

The purpose of ChIP-seq and ChIP-chip is to use the distribution of the immunoprecipitated DNA within the genome to locate positions where proteins bind to the DNA and also assess the degree of binding, which is done by looking at the locations and sizes of the peaks in the distribution.

Unfortunately, there are frequently peaks in the distribution of the immunoprecipitated DNA that do not correspond to protein binding locations but are instead artefacts that arise out of the bio-chemistry of the process and also the bioinformatics of the alignment process. One of the ways of demonstrating that these are artefacts is to examine the fragment distribution in a control which does not involve the immunoprecipitation of DNA fragments attached to the target protein. If the peaks are also present in the control then their presence in the immunoprecipitated fragments is assumed to be as a result of an artefact in the process and the peaks are ignored [52]. There are two ways in which a control can be generated for such experiments [77 p 672].

The first approach is to perform a ‘mock’ immunoprecipitation of a sample of the fragments using an antibody that selects for a protein that is not believed to be present in the sample. However, it has been found that when this is done, the quantities of DNA fragments are often so low that the information that is derived from this control is of very poor quality. One possible solution to this problem is to add more stages of PCR amplification for these fragments, but that introduces the risk of making any slight contamination of the sample more significant. It also makes the sample less useful as a control in that there are additional differences in the way that the two samples have been treated.

A second approach that has been adopted in order to avoid the problems of low sample quantities from mock immunoprecipitation is to use a sample of the input DNA as a control, as it has been shown that many of the artefacts also show up as peaks in the input DNA, providing the information needed to remove the artefacts in the sample

In either case, it is common practise not to process a control that is associated with every set of immunoprecipitated fragments, but instead to produce a control for a particular cell line that has been treated in a particular way, and use it as the control for a large number of experiments that have been done with the same cell line and treatment. This is done on the assumptions that there is not a significant variation in the distribution that is seen between controls and that the artefacts that are seen are common to all the controls and so it is not necessary to produce a separate control for each experiment.

1.4.6 Fragment quantification using ChIP-chip

ChIP-chip is the earliest of the two technologies and uses microarrays to quantify the amount of DNA from different regions of the genome [63]. Microarrays are small glass slides that have been spotted with 100's of thousands of individual spots, each containing thousands of identical copies of a short sequence of single stranded DNA, typically between 30 and 50 nucleotides long. Each spot contains DNA with a different sequence, and each will match a specific location within the genome. When the sample is washed over the array, fragments whose sequence matches the sequence of the DNA on a specific spot will bind, and the overall quantity bound at any spot can be measured using fluorescence.

One of the drawbacks with this approach is that it is only possible to quantify the DNA at a limited number of locations in the genome. It is therefore necessary to decide in advance which locations are likely to be of interest and construct a microarray that is engineered around the sampling choice that has been made. It is frequently the case that the array will only provide one or two probes for each gene, although chips are available that have four probes for each exon and consequently roughly forty probes for each gene. It is therefore quite possible that significant quantities of fragments, corresponding to a location in the genome that does not have an associated probe perhaps through incomplete knowledge of the way that DNA is transcribed, could be missed.

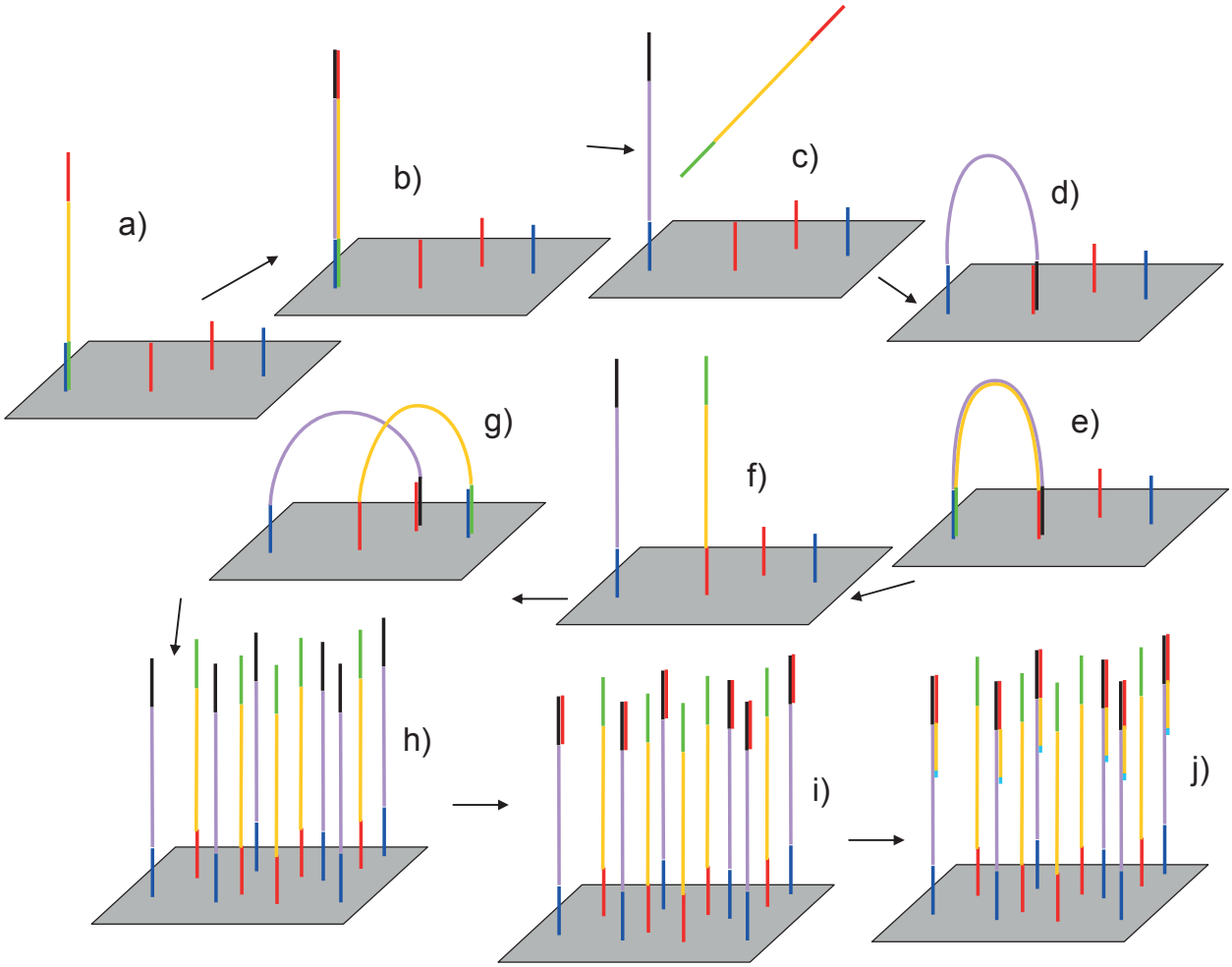


Figure 1-4 Amplification and sequencing of fragments on the flow cell. a) After amplification the DNA is denatured and run over the flow cell which is covered with a lawn of primers (red and blue) corresponding to the sequence of one the two adaptors and the complement of the other. Fragments (orange) bind at relatively widely dispersed locations, and will only bind at one end as only the green or black adaptor sequences matches a lawn primer. b) DNA polymerase is used to complete the second strand (lilac), starting with the primer that is attached to the flow cell and proceeding in a 5' to 3' direction. c) The strands are denatured, and the original strand is now lost as there are no complementary primers on the flow cell to which it can bind. d) The end of the new strand binds to a nearby complementary primer on the flowcell lawn. e) DNA polymerase is used to complete the second strand, again starting with the bound primer. f) The strands are denatured leaving two complementary strands attached to the flow cell g) The ends of the strands bind to two further primers h) The process continues until the spot consists of a mix of equal quantities of the original strand and its complement. i) Primers are added for one of the two free adaptors. j) The second strand for the half of the strands are synthesised one nucleotide at a time in the 5' to 3' direction using fluorescent nucleotides, the last one that has been added is shown in light blue. For the strands where the synthesis remains synchronized, the additional nucleotides will be identical, and their combined fluorescence allows the identity of the nucleotide to be determined.

1.4.7 Fragment identification and quantification using ChIP-seq

This protocol builds on the experience the ChIP-chip protocol but locates the fragments on the genome using the fast developing, high throughput sequencing technology. This removes the requirement to decide in advance what regions will be sampled and which excluded, and it removes the need to create chips with specific sequences for each organism, and chips for each sampling strategy associated with an organism. There are extensive reviews of both technologies, and comparisons between the two, in the literature [e.g. 30].

In order to locate the source of each fragment within the genome, the end of each fragment is sequenced, resulting in a sequence tag, and the sequence then located within the genome. There are a number of variants of the sequencing protocols that are used to sequence DNA and which can be used within the ChIP-seq process. They do however share many underlying principles. The details of the process employed by the Illumina sequencing platforms are shown in Figure 1-4.

For any given sequence length there will be a certain proportion of sequences that do not map to a unique location in the genome. A number of approaches can be adopted for dealing with such fragments, but the commonest is to disregard them, and chose a sequence length that is sufficient to keep the proportion of disregarded fragments to an acceptably low level. Sequence lengths of 25 or 36 are commonly chosen and have been found to be adequate for sequence alignment to the human genome.

1.4.8 Definition of fragment orientation

Any given sequence tag may align to either the forward strand or the reverse strand of the DNA. The double stranded sequence is molecularly symmetrical, and the choice of which strand is the forward strand was an arbitrary necessity for the purposes of nucleotide numbering. When the fragment is aligned to the genome the convention adopted in this thesis is to define the fragment start and end with respect to the forward strand direction. A sequence tag that aligns to the forward strand thus identifies a fragment start, and a tag alignment to the reverse strand identifies a fragment end (Figure 1-5a).

At other times, the relationship of the fragment to the local sequence, particularly a sequence that matches a known protein binding sequence, is more important in which case the fragment direction is defined with respect to the direction of this sequence (Figure 1-5b). Again, there can be an arbitrariness about this as the initial choice of a sequence or its reverse complement as the direction of the reference may be arbitrary. One exception to this is if the

sequence is always in a specific orientation with respect to the protein coding sequence, whose direction of transcription can then define the orientation of the sequence.

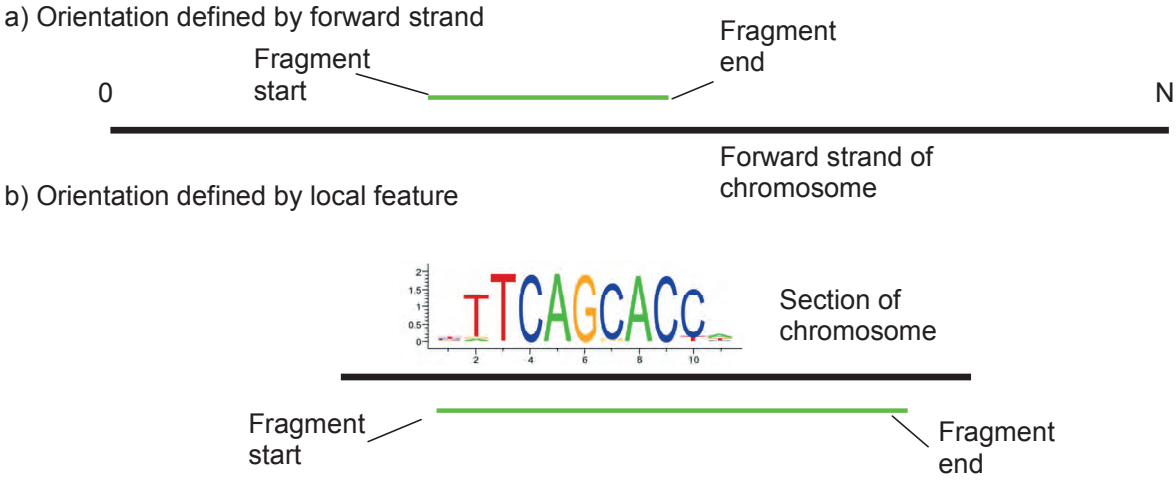


Figure 1-5 Definition of fragment start and end. Green fragments are shown aligned to a region of the genome (black) a) Fragment direction defined with respect to the forward strand of a chromosome. b) Direction defined with respect to some local feature such as a nucleotide sequence that matches a specific pattern on a specific strand.

1.4.9 ChIP-seq peak finding algorithms

In the case of ChIP-seq, the fragment location information is then collated. An example of a typical fragment distribution in the region of a transcription factor binding site was shown in Figure 1-1. Various algorithms and their implementations in software have been developed to interpret these data in order to identify the probable locations of the transcription factor binding sites that gave rise to the ChIP-seq results that were obtained [94]. A common factor to all of these algorithms is that they use a smoothed version of the fragment distribution, often by a summation of the counts over a specific window size such as 25 nucleotides. One advantage of such windowing is that it makes the computation more tractable. There is an implicit assumption in doing this that there is no significant information associated with the individual counts, and the nucleotide to nucleotide variation in count is essentially noise that arises out of the fragmentation process..

The simplest approach to identifying the locations where the proteins were bound is to identify the regions with significant numbers of fragments, a technique that was used by some of the earliest peak finders [49, 83]. Subsequent algorithms have made more use of the more detailed information that is available from the data. This includes making use of the different

distributions of the forward reads and the reverse reads such as is used in cisGenome [46] or in the QuEST software [96]. A common feature of many peak finders is that the distribution of the immunoprecipitated fragments is compared to the distribution of the control sample such as the input DNA [47, 85] in order to remove peaks in the data that are as a result of artefacts in the process and not protein binding (Section 1.4.5). These algorithms look for the degree to which the fragments are enriched in some regions as a result of immunoprecipitation compared to the control. They ignore those peaks that are present in the immunoprecipitated DNA because there was a corresponding peak in the input fragment distribution and no enrichment had occurred as a result of immunoprecipitation. Such peaks in the input fragments are as a result of other factors such as the effect of DNA structure on fragmentation or alignment artefacts such as those shown in Figure 1-6.

Bayesian techniques have also been used to interpret the data [91] and other software has been developed to looking specifically for the characteristic shapes of the peaks that indicate a binding site [44]. Examples of such peaks can be seen in Figure 2-17.

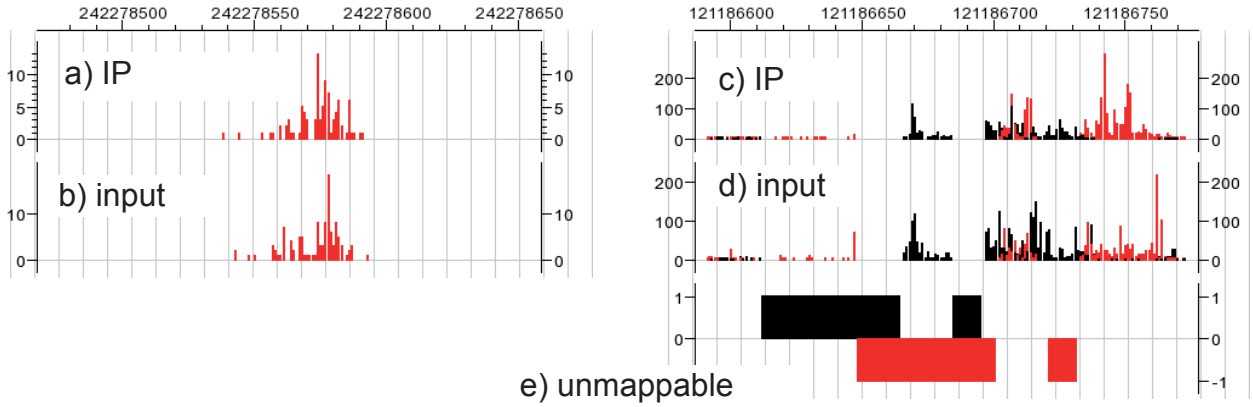


Figure 1-6 Examples of artefacts where peaks are seen in both input and immunoprecipitated tags. a) & c) Peaks in tag distributions of immunoprecipitated DNA where peaks are also seen the input DNA (b & d). e) Unmappable regions. All plots show a region from chromosome 1 and black indicates forward strand and red the reverse strand. c) & d) are characteristic of artefacts frequently found immediately adjacent to unmappable regions and suggest that fragments from very repetitive regions of the genome that have yet to be sequenced are mapped to other similar slightly less repetitive regions which have been sequenced. They have been mapped because there is only one instance of the specific tag sequence in the sequenced genome.

1.4.10 Motif finding

Once peak finding has been used to identify regions where the target transcription factor was bound, a common task is then to identify the specific locations within these regions

where the protein was bound. One general principle that is frequently adopted is to look for a DNA sequence that is over-represented within these regions compared to the frequency with which it occurs within the overall genome. One improvement that can be made is make the comparison with carefully selected control regions of the genome that are similar in some respect to the region where the peak is located but where was no significant degree of protein binding.

One difficulty that is encountered in motif finding is that transcription factors often do not bind to one specific DNA sequence but instead bind to a range of sequences that are variants of some underlying pattern. The degree of binding is a complex function of the DNA sequence in combination with many other factors such as chromatin conformation and the binding of other proteins nearby on the DNA.

Many algorithms have been developed to try and identify the sequence pattern or motif associated with protein binding. One well established and frequently used algorithm is MEME (Multiple EM for Motif Elicitation) [5]. This built on an early Expectation Maximization (EM) algorithm [58] which assumed that there was a single instance of a variant of the MOTIF in each of the regions and then finds the sets of positions *i* in the regions *j* that gives the maximum likelihood that they are the sequences derived from an underlying motif.

The first extension added by MEME was to use sequences from the regions as initial seeds for searching for the underlying motif. Another extension was to remove the assumption that there is only one instance of the sequence in each region. The final extension is that once a motif is found the sites associated with the motif are deemphasised during the search for additional motifs, which improves the ability to find multiple alternative binding motifs.

The motif finding algorithm that is integrated into cisGenome [47] and which is used for motif dinding in this research uses a Gibbs sampling algorithm and a Bayesian approach to identifying conserved motifs within the regions of the ChIP-seq peaks [45, 64]. In common with MEME, it is able to detect multiple alternative motifs, and also to detect motifs that occur more than once within any one of the set of regions being examined.

1.4.11 Representation of motifs using Position Specific Scoring Matrices (PSSMs)

The simplest way of representing the overrepresented motifs is with a consensus sequence which identifies the nucleotide that is most frequently found at each position within the binding region. However, such an approach is not subtle enough to capture the characteristics of the binding motif for many proteins. These can usually bind when some

nucleotides do not correspond to the consensus sequence, and where the motif is better expressed in terms of a tendency towards certain nucleotides at certain positions.

The approach frequently adopted is to represent the motif as a Position Specific Scoring Matrix (PSSM) or Position Weight Matrix (PWM) **P** such that

$$\begin{aligned} \mathbf{P} &= \left(\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3 .. \mathbf{p}_N\right) \\ \mathbf{p}_i &= \left(w_{i,\mathrm{a}}, w_{i,\mathrm{c}}, w_{i,\mathrm{g}}, w_{i,\mathrm{t}}\right) \end{aligned} \tag{1.1}$$

Each member $\mathbf{p}_i$ of the ordered set **P** is a vector of weights $w_{i,n}$ which are a measure of the probability of finding the nucleotide n at position i. These can be derived by experimentally identifying locations in the DNA where the protein is known to bind, and counting the number of each of the nucleotides at each of the positions and using the counts as the values $w_{i,n}$. If M positions have been identified then

$$\sum_{n \in \{\mathrm{a,c,g,t}\}} w_{i,n} = M \tag{1.2}$$

It is frequently the case that the values are normalised such that

$$\sum_{n \in \{\mathrm{a,c,g,t}\}} w_{i,n} = 1 \tag{1.3}$$

When this is done, then the weights give the likelihood of a nucleotide being found at a particular position if it is known to be a location where the protein binds.

This information can then be used to determine the likelihood of any given nucleotide sequence being associated with a particular motif. For each position in the sequence the likelihood of the nucleotide is compared with the likelihood given the background nucleotide position, and the overall likelihood can then be calculated as the product of the likelihood at each position.

The background distribution that is used can be a simple independent likelihood of each nucleotide, or can be a more complex likelihood, for example based on a third order Markov model of the nucleotide sequence characteristic that is derived for each region of the genome. Such a Markov model is used in motif mapping software such as is found in the cisGenome software suite [46].

It is also frequently the case that the log likelihood is used rather than the likelihood. This allows the log likelihood to be calculated from the sum of the log likelihoods for each nucleotide position within the PSSM.

1.4.12 The advantages and disadvantages of ChIP-chip

ChIP-chip and ChIP-seq are both widely used technologies for determining the level of protein binding to DNA. ChIP-chip is the earliest of the technologies and is still the cheapest and offers the possibility of having results in a matter of days. Its disadvantage is that it is essentially an analogue technology which limits the precision of the measurements, and the results are spatially quantised along the genome, as set by the spacing along the genome of the probes that are used to bind to the DNA. Consequently it only allows the measurement of DNA levels at a specific set of points along the genome. The more recent ChIP-seq process [77], building on the capabilities of high throughput sequencing technologies, has the advantage that it has a greater dynamic range, being able to measure differential protein binding over many orders of magnitude, and can also provide very detailed binding information, down to the level of individual nucleotide positions (Chapter 4). It is nevertheless the case that it is relatively expensive and has a long turnaround time from starting the experiment to obtaining final results.

1.4.13 The use of sonication to explore chromatin structure

There has already been some recognition that the locations where DNA breaks during the sonication stage of the ChIP-seq protocol are not uniformly distributed, and can in themselves be used to investigate genomic characteristics. It has been found [3, 95] that fragments send to occur preferentially in the regions where the chromatin is more open, such as is the case in active promoter regions. These investigations also showed that the degree to which the fragments were able to identify the open regions of the chromatin was dependent on the size of the fragment that was selected after sonication, with the shorter fragments being better indicators of open chromatin [3].

1.5 An introduction to the RNA-seq protocol

The use of high throughput sequencing techniques for analysing RNA data has provided another way in which sequencing can be used to probe the mechanisms within the cell that enable the genetic code to be involved in the control of almost every aspect of the functioning of a cell.

There are a variety of different types of RNA in the cell, and different variants of RNA sequencing protocols have been developed that are tailored to each type of RNA species. For example, the growing interest in the role of microRNAs[80], has resulted in the increased use of high throughput sequencing technologies for quantifying the types and levels of expression

of these microRNAs [31]. The lengths of these RNAs mean that the fragmentation stage that is necessary in DNA sequencing is not required, as each microRNA is short enough to be sequenced in its entirety.

However, the first and still the dominant RNA based protocol is RNA-seq, where the RNA is extracted from the cell, purified and then fragmented so that, by sequencing the ends of the fragments that are obtained, a picture can be constructed of the nucleotide sequence of the RNA molecules that are found in the cell and their relative abundance. The primary purpose of this procedure is to investigate messenger RNA (mRNA) in the cell, where it can be used to investigate the expression levels of genes, and to investigate the different transcript variants [73, 98]. It is then able to be used to refine the genome annotation to provide more detail of the differential transcript expression.

RNA sequencing protocols contain many of the elements from the ChIP-seq like protocols, but there are a number of very specific and important differences. In particular, high throughput sequencing protocols have been designed around the processing and sequencing of DNA, and many of the stages of the procedure such as the amplification using PCR are only currently possible using DNA. Consequently a necessary step in the sequencing of RNA fragments is the conversion of RNA into the complementary DNA sequence using the reverse transcriptase enzyme so that it can then be amplified and sequenced [98]. Reverse transcriptase can only initiate the transcription at a location in the RNA where the complementary strand of DNA is already present which acts as a primer for the process.

The problem that has to be solved when transcribing the RNA fragments to DNA is that the sequences of the fragments are all very different, making it difficult to design a primer or set of primers which will bind to the RNA and act as a primer and allow all of the fragments to be converted to DNA in an unbiased way.

The solution that has been widely adopted is to create a set of DNA primers that are six nucleotides long and are a random mix of the 4096 different combinations of six nucleotides that are present in such hexamers [32][1]. A six nucleotide length primer is sufficiently long to enable reverse transcriptase to bind and start the reverse transcription process, and there will be a hexamer present in the mix that can bind to any position in an RNA fragment.

[1] The primary subject of this journal article was the hypomethylation of cancer genes. However this article now has almost 1000 citations which are almost entirely to the random primer method that was developed for the experiment and which is incidental to the main purpose of the article.

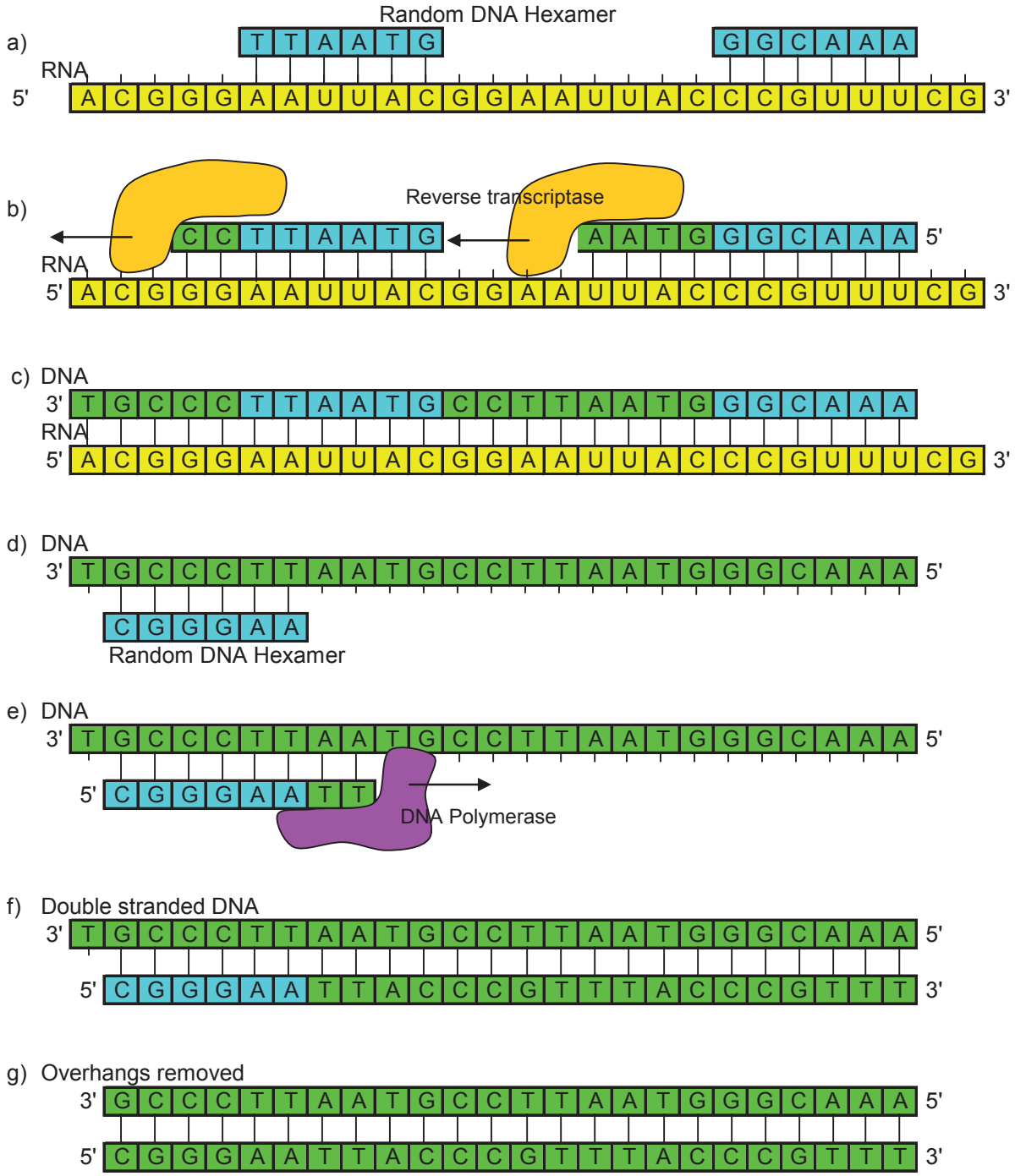


Figure 1-7 RNA fragment is converted to a slightly shorter double stranded DNA fragment during reverse transcription. a) Random DNA hexamers binds to RNA. b) Reverse transcriptase completes a complementary DNA strand c) Completed first DNA strand d) RNA removed with RNase and random hexamer binds to DNA. e) DNA polymerase completes second strand. f) Second strand completed g) Overhangs removed with T4 DNApolymerase and Klenow DNA polymerase. The DNA fragment is slightly shorter than the original RNA, the degree of shortening being determined by the positions where the hexamers bind.

The conversion is a two stage process. In the first stage, complementary DNA is added to the single stranded RNA fragments in order to create a double stranded polynucleotide that is a pairing of DNA and RNA. These are then separated, and the single stranded DNA converted to double stranded DNA using DNA polymerase (Figure 1-7).

Both of these processes require a DNA primer to be bound to the DNA or RNA to allow the enzyme to bind and reverse transcription or polymerisation to proceed. In both cases the primers will bind at multiple locations, and the transcription will proceed at each location until the enzyme meets the location of the next primer. At this point the enzyme can complete the second strand right up to the primer. In both cases however, transcription can only start from the position where the first primer bound, and this may not be right at the start of the template RNA or DNA fragment. A consequence of this is that the ends of the final DNA fragment that is sequenced will not correspond to the ends of the fragment that was originally formed when the RNA was fragmented.

1.6 Introduction to the thesis

One feature of the use of next generation sequencing in the ChIP-seq protocol is that while each experiment generates very large quantities of data, much of this is discarded during the processing of the data to identify protein binding sites.

The area of research that was the original motivation for the work described in this thesis was a suspicion that within this discarded data there were data that could be used to provide more information about the nature of the binding of transcription factors to DNA. The initial investigation centred on the pattern of fragment starts in ChIP-seq data that is associated with over-represented sequences or motifs in the vicinity of peaks in the ChIP-seq data. The initial conclusion is that such patterns can provide an additional source of information about the way that proteins such as transcription factors bind to these sequences. This is covered in Chapter 4.

The reason why this is covered in one of the later chapters of this thesis is that it quickly became clear that there were other genome wide sequence specific effects that influence the probability of the DNA fracturing at any specific location in the genome. It was thought that these effects would need to be analysed further in order to be able to compensate for them if necessary and so get a clearer picture of the way in which proteins that are bound to the DNA might influence DNA fragmentation.

The major effect that was investigated was a genome-wide bias in relationship between fragment start site locations and the DNA sequence in the immediate vicinity of these

locations. It was clear that fragment starts were associated significantly more frequently with some sequences and less frequently with others. While a degree of sequence dependency has previously been identified, the details, origin and potential impact on the interpretation of results are still poorly understood [25, 26, 88].

This investigation is covered in Chapter 2 which describes a new modelling technique which shows how this sequence dependency is significantly more complex than previously realised. For example, there is a significant variation in this effect between different experiments, which was previously unrecognised. The chapter also describes the mathematical model that was developed to describe the relationship between DNA sequence and the probability of fragmentation at any specific location, and the model fitting algorithm that was used to fit the model parameters to the observed data. The more detailed results obtained from this modelling technique allows more informed discussion on the origin and impact of this dependency.

After this model had been developed it was realised that it could also be used to provide a better picture of the sequence bias that occurs in the RNA-seq protocols. This investigation is covered in chapter 3 which mirrors the contents of chapter 2 in that it shows that the effect is significantly more complex than previously acknowledged and that in this complexity can be found information that can help understand the mechanisms that occur during the RNA-seq process.

When the ends of fragments are sequenced in order to align them to the genome, the finite read length means that there will be some sequences where it will not be possible to identify a unique location where the fragment originated in the genome. In this situation the simplest and most commonly adopted practise is to ignore these fragments. This means there will be some regions of the genome where the fragment start density is zero. This is an artefact that needs to be corrected for in order to improve the accuracy of the analysis of sequence dependent fragmentation covered in Chapter 2. At the time of starting this analysis there were no generally available tools for locating the regions in the genome that are unmappable in this way, which is a necessary prerequisite to being able to compensate for this artefact. Consequently an algorithm and associated software was developed that allowed these regions to be identified efficiently. This is described in Appendix A.

Chapter 5 then describes how the model fitting approach and algorithms that were created to identify sequence dependent characteristics in high throughput sequencing data were also suitable to be applied to extract additional information from ChIP-chip data. The

chapter describes a specific example where this approach was used to obtain greater detail about the way that the SeqA protein binds *in vivo* to the *E. coli* genome.

Chapter 2

Sequence bias in ChIP-seq experiments

This chapter presents a new model-fitting approach to the analysis of the sequence bias at the start of DNA fragments in ChIP-seq experiments.

2.1 Introduction

One of the underlying assumptions in using the ChIP-seq protocol (Section 1.3) to examine the distribution of bound proteins throughout the genome is that the DNA fragmentation, which is a key stage in this protocol, is either random or is determined primarily by the characteristic being investigated and is otherwise independent of other genomic features such as the background genomic sequence. Unfortunately, bias has been observed to occur when DNA is fragmented, raising concerns that such bias will influence any conclusions drawn from these data [25, 26, 88].

One potential source of bias that has been investigated is sequence dependent misreads that can occur in DNA sequencing [25]. A further source of bias in ChIP-seq data has been found to be associated with regional GC content [88].

Some evidence for bias in the nucleotide sequences at the start of the fragments from ChIP-seq experiments has been found [88]. As well as investigating the bias in order to improve the protocols, the knowledge gained about the bias has been used to develop ways of manipulating the data to reduce the effect of the bias [16, 40, 60, 88].

In all the previous work the underlying assumption has been that there is a single nucleotide sequence pattern that describes the bias observed at the start of the immunoprecipitated fragments, and that very similar characteristics are seen when the results of different experiments are compared.

Here we demonstrate that for DNA fragmentation the sequence characteristics are more complex than can be modelled using a single sequence pattern, and that modelling based on multiple alternative sequence patterns within a single experiment provides a considerably richer and more detailed picture of what happens during the experimental procedure. It is then possible to better understand the differences in bias between experiments that have been observed [88]. It is also possible to identify situations where the effect of bias has been hidden as a result of using simple models that had the effect of averaging out a complex underlying picture.

The analysis also demonstrates significant bias in the nucleotides immediately preceding the start of the fragment, a possibility that is ignored in other analyses of bias.

2.1.1 Definition of sequence bias

There are two different but related definitions that can be used for sequence bias. The definition that has often previously been used [e.g, 88] is based on looking at the distribution of the nucleotides at the start of the DNA fragments that are ultimately sequenced. This is $p(\mathbf{s}|\theta)$, the probability in a particular ChIP-seq dataset of finding a sequence **s** given a fragment θ and is directly measureable from the fragment data. One implication of using this approach is that it is influenced by the relative numbers of each nucleotide sequence **s** in the genome. An under- or overrepresentation of specific sequences in the genome will result in an under or overrepresentation of these sequences at the start of fragments, even if the fragments are uniformly or randomly distributed in the genome in a way that is not dependent on the sequence.

An alternative approach is to consider the probability that any fragment selected from the dataset will start at a particular position in the genome given the local sequence **s**, i.e. $p(\theta|\mathbf{s})$. If the fragmentation is sequence independent then this probability will be equal for all values of **s.** The two ways of calculating $p(\theta,\mathbf{s})$, the joint probability of a finding a fragment with a sequence **s** (2.1), can then be used to determine the relationship between the two bias definitions.

$$p(\theta,\mathbf{s}) = p(\theta|\mathbf{s})\,p(\mathbf{s}) = p(\mathbf{s}|\theta)\,p(\theta) \tag{2.1}$$

$p(\theta)$ is the sequence independent probability of a fragment coming from a specific location, which can be calculated by assuming the fragment has equal probability of occurring at any location. $p(\mathbf{s})$ is the probability of finding a specific sequence in the genome and is the measure of the variation in the number of occurrences of each of the sequences within the genome.

It is the probability $p(\theta|\mathbf{s})$ that has been used in this thesis when looking at ChIP-seq fragmentation. This is because it is a measure of how much the local sequence influences the probability that the DNA will fragment at that point, defined such that it is independent of the specific sequence distribution in the genome. In order to make comparisons easier, the

probability $p(\theta|\mathbf{s})$ is normalised to $Y_{\mathbf{s}}$ which equals one if the probability of a fragment is sequence independent. This can be done by dividing $p(\theta|\mathbf{s})$ by $p(\theta)$.

$$Y_s = \frac{p(\theta|\mathbf{s})}{p(\theta)} \tag{2.2}$$

Another way of thinking of $Y_{\mathbf{s}}$ is as the ratio of the probability of finding a fragment with a specific sequence where there is a joint dependency and the probability of finding a fragment with a specific sequence if there is no dependence (2.3). This will tend to one if $p(\theta)$ and $p(\mathbf{s})$ are independent and is derived from (2.2) as follows:

$$\begin{aligned} Y_s &= \frac{p(\theta|\mathbf{s})}{p(\theta)} \\ &= \frac{p(\theta|\mathbf{s})\,p(\mathbf{s})}{p(\theta)\,p(\mathbf{s})} \\ &= \frac{p(\theta,\mathbf{s})}{p(\theta)\,p(\mathbf{s})} \end{aligned} \tag{2.3}$$

Section 2.2.4 describes how a weighted summation of a similar ratio can be used to calculate the mutual information between the DNA sequence at a specific location and the probability of fragmentation.

While $p(\theta|\mathbf{s})$ is not directly measureable, a definition of $Y_{\mathbf{s}}$ in terms of measureable data can be derived by using (2.1) as follows:

$$\begin{aligned} Y_s &= \frac{p(\theta|\mathbf{s})}{p(\theta)} \\ &= \frac{p(\mathbf{s}|\theta)\,p(\theta)}{p(\mathbf{s})} \times \frac{1}{p(\theta)} \\ &= \frac{p(\mathbf{s}|\theta)}{p(\mathbf{s})} \end{aligned} \tag{2.4}$$

For simplicity $Y_{\mathbf{s}}$ is referred to as sequence bias in this document, even though this is not identical with the usage of this term in other documents where it is related to $p(\mathbf{s}|\theta)$, the probability of finding certain sequences or nucleotides associated with fragments. The two definitions give identical values if there is an equal number of all of the different sequences being considered within the genome.

2.2 Method

2.2.1 Summary of data sources †‡

For most of this investigation, data from the ChIP-seq input fragments were used in order to avoid possible sequence bias that might arise from using immunoprecipitated fragments where sequences are preferentially drawn from regions where the target protein binds (see Section 1.4.5).

ChIP-seq data will normally contain a mixture of fragments that originate in the nuclear DNA and the mitochondrial DNA. The mitochondrial genome is significantly smaller than the nuclear genome, but is present in a much higher copy number within the cell. The average fragment density $p(\theta)$ is therefore different for the nuclear and mitochondrial DNA. If both sorts of DNA are used in the calculations then in the definition of Y_s in (2.4) the $p(\mathbf{s})$ factor for such data is no longer simply a function of the number of each N-mer in the genome, but is now a more complex function where the mitochondrial and nuclear DNA are treated separately, weighting each component by the relative concentrations of the two types of DNA. Rather than introducing this additional complexity into the definitions, the mitochondrial DNA has simply been excluded from the analysis.

The analysis was performed using 12 sets of input DNA from *Homo sapiens* ChIP-seq experiments conducted by the Myers/HudsonAlpha lab [49, 96], 11 sets of input DNA from the *Homo sapiens* ChIP-seq experiments conducted by the Yale/UCD/Harvard labs, 2 *Homo sapiens* datasets published as part of an investigation into the mapping of HATs and HDACs [99] , a set of 4 input data from *Caenorhabditis elegans* ChIP-seq experiments [20] and a set of data from *Arabidopsis thaliana* produced at the University of Warwick.

The following provides more details of these data sources.

Data from the Myers/HudsonAlpha lab,

Input fragment data from ChIP-seq experiments on various *Homo sapiens* cell lines and types which had been produced as part of the Encyclopaedia of DNA Elements (ENCODE) project. These were obtained from:

http://hgdownload.cse.ucsc.edu/goldenPath/hg18/encodeDCC/wgEncodeHudsonalphaChipSeq/

Lab Version	File	Cell line	Protocol (a)	Treat	anti body	Replicate	GC-rich (b)
SL116	wgEncodeHudsonalphaChipSeqAlignmentsRep1Panc1Nrsf.tagAlign.gz	PANC1	PCR2x	None	NRSF	1	Y
SL117	wgEncodeHudsonalphaChipSeqAlignmentsRep1Panc1Control.tagAlign.gz	PANC1	PCR2x	None	input	1	Y
SL522	wgEncodeHudsonalphaChipSeqAlignmentsRep2Panc1Nrsf.tagAlign.gz	PANC1	PCR2x	None	NRSF	2	N
SL523	wgEncodeHudsonalphaChipSeqAlignmentsRep2Panc1Control.tagAlign.gz	PANC1	PCR2x	None	input	2	N
SL102	wgEncodeHudsonalphaChipSeqAlignmentsSknmcControl.tagAlign.gz	SK-N-MC	PCR1x	None	Input	2	Y
SL103	wgEncodeHudsonalphaChipSeqAlignmentsRep1U87Control.tagAlign.gz	U87	PCR2x	None	Input	1	Y
SL217	wgEncodeHudsonalphaChipSeqAlignmentsRep1Gm12878ControlPcr2x.tagAlign.gz	GM12878	PCR2x	None	input	1	Y
SL218	wgEncodeHudsonalphaChipSeqAlignmentsRep2Gm12878ControlPcr2x.tagAlign.gz	GM12878	PCR2x	None	input	2	Y
SL516	wgEncodeHudsonalphaChipSeqAlignmentsRep1Gm12878ControlV2.tagAlign.gz	GM12878	PCR1x	None	input	1	N
SL517	wgEncodeHudsonalphaChipSeqAlignmentsRep2Gm12878ControlV2.tagAlign.gz	GM12878	PCR1x	None	input	1	N
SL518	wgEncodeHudsonalphaChipSeqAlignmentsRep1K562ControlV2.tagAlign.gz	K562	PCR1x	None	input	1	N
SL519	wgEncodeHudsonalphaChipSeqAlignmentsRep2K562ControlV2.tagAlign.gz	K562	PCR1x	None	input	2	N

(a) The data were produced using two different amplification methods, as designated in the table:
PCR2x: Two rounds of amplification, 25 and 15 cycles
PCR1x: One round of amplification, 15 cycles

(b) "GC-rich" is an indication as to whether or not the bias at the fragment end conformed to the GC-rich pattern shown by SL117 or the more varied pattern similar to that shown by SL523 (Section 2.3.6).

The Myers lab used different sonication methods during the period covered these experiments. The following excerpt from the protocol description used by the Myers/HudsonAlpha lab provides more details [78].

"Note: The Myers lab has used two different methods for sonicating chromatin. All of our experiments until Fall 2009 used a Sonics VibraCell sonicator, a relatively inexpensive approach that we fine-tuned to fragment the chromatin to a specific size range. After that time, we began using a Bioruptor sonicator, which is much easier (multiple samples can be sonicated at the same time) and cleaner (the samples are closed during the sonication treatment). The reagents used are the same, but the methods differ."

Data from the Snyder/Yale lab

Input fragment data from ChIP-seq experiments on various *H. sapiens* cell lines and types produced as part of the ENCODE project obtained from:

http://hgdownload.cse.ucsc.edu/goldenPath/hg18/encodeDCC/wgEncodeYaleChIPseq/

Name	File	Cell	Treat	antibody	Replicate
Y633-1	wgEncodeYaleChIPseqAlignmentsRep1K562InputV3.tagAlign.gz	K562	None	input	1
Y633-2	wgEncodeYaleChIPseqAlignmentsRep2K562InputV3.tagAlign.gz	K562	None	input	2
Y787-1	wgEncodeYaleChIPseqAlignmentsRep1Helas3MouseiggV2.tagAlign.gz	HeLa-S3	None	input	1
Y787-2	wgEncodeYaleChIPseqAlignmentsRep2Helas3MouseiggV2.tagAlign.gz	HeLa-S3	None	input	2
Y864-1	wgEncodeYaleChIPseqAlignmentsRep1K562MusiggMusigg.tagAlign.gz	K562	None	input	1
Y864-2	wgEncodeYaleChIPseqAlignmentsRep2K562MusiggMusigg.tagAlign.gz	K562	None	input	2
Y956-1	wgEncodeYaleChIPseqAlignmentsRep1Gm12878MusiggMusigg.tagAlign.gz (GM12878_IgG_Control_tagAlign_rep1_FC30P42HM_20081212_s_6.)	GM12878	None	input	1
Y1066-1	wgEncodeYaleChIPseqAlignmentsRep1Hepg2ControlForskln.tagAlign.gz	HepG2	forskolin	input	1
Y1066-2	wgEncodeYaleChIPseqAlignmentsRep2Hepg2ControlForskln.tagAlign.gz	HepG2	forskolin	input	2
Y1109-1	wgEncodeYaleChIPseqAlignmentsRep1Gm12878InputIggrab.tagAlign.gz GM12878_Rabbit_IgG_tagAlign_rep1_100106_ROCKFORD_FC600AF_s_4	GM12878	None	input	1
Y1109-2	wgEncodeYaleChIPseqAlignmentsRep2Gm12878InputIggrab.tagAlign.gz GM12878_Rabbit_IgG_tagAlign_rep2_100107_COLUMBO_FC600AU_s_4	GM12878	None	input	2

Data previously analysed by Wang et al [99]

Data on various *H. sapiens* cell lines and types obtained from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) database [6].

GSM	File	Cell	Treat	Antibody	Replicate
GSM393947	GSM393947_CD4-PCAF.bed.gz	CD4+ T cell	None	PCAF	1
GSM418301	GSM418301_HeLa-siControl-H3K9ac-HDACi-0h.bed.gz	HeLa	None	None	-

Data previously analysed by Cheung et al [20]

Data from various *C. elegans* ChIP-seq experiments obtained from the GEO database [6]. Raw sequence data extracted from the sra file, and aligned to the UCSC version 6 of the *C. elegans* genome based on Wormbase WS190 using the -m 1 option so that sequences that map to multiple locations are excluded.

GSM	File	Strain	Stage
GSM706161	SRR190662.sra	N2	L3
GSM706164	SRR192330.sra	N2	L3
GSM727910	SRR210889.sra	N2	L3
GSM727911	SRR210890.sra	N2	L3

Arabidopsis thaliana input DNA

This data is the input DNA for an LHY chip-seq experiment conducted at Warwick University.

2.2.2 Definition of ChIP-seq sequence bias ‡

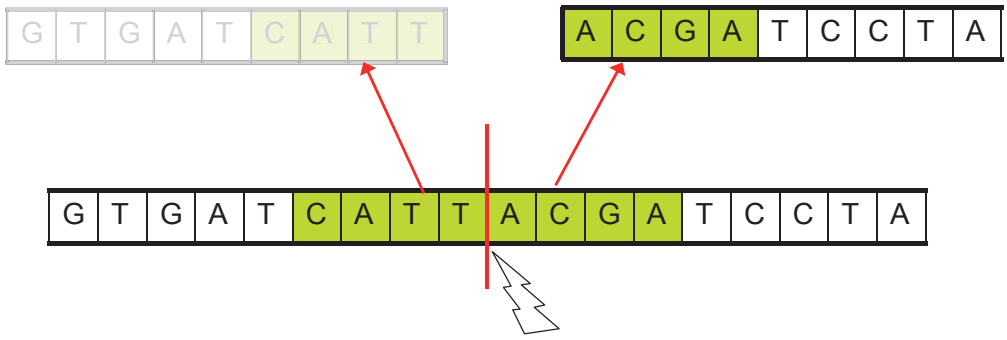


Figure 2-1 Relationship between the probability of DNA fragmenting and the local DNA sequence. The bias is the extent to which the sequence such as the 8-mer sequence marked in green determines the probability of DNA fragmentation at a specific position within the 8-mer. In this example it is the influence of the sequence on the likelihood of fragmentation at the midpoint that is being considered which in this case creates a fragment starting 'ACGA'.

The approach adopted for analysing sequence bias was to calculate how much the DNA sequence affects the probability of a fragment starting at a specific location. Consider an 8-mer $\mathbf{s}$ (e,g, CATTACGA in Figure 2-1) that occurs $N_{\mathbf{s}}$ times within the genome and let $C_{\mathbf{s}}$ be the number of fragments that start at a specific position within the 8-mer (e.g. the midpoint in the Figure 2-1).

From (2.4) two probabilities are required, $p(\mathbf{s}|\theta)$ and $p(\mathbf{s})$. These can be derived from measureable data as follows:

$$p(\mathbf{s}|\theta) = \frac{C_{\mathbf{s}}}{C_{tot}}$$
$$p(\mathbf{s}) = \frac{N_{\mathbf{s}}}{G_r} \qquad (2.5)$$

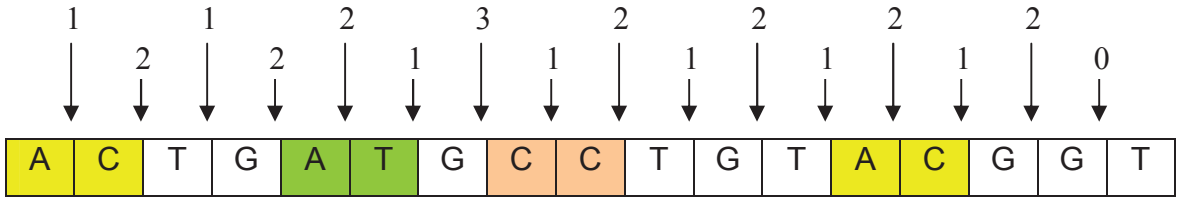
$C_{\mathbf{s}}$ is the number of fragments associated with the 8-mer $\mathbf{s}$ in the dataset and C_{tot} is the total number of aligned fragments in the dataset. $N_{\mathbf{s}}$ is the total number of the 8-mer $\mathbf{s}$ in the genome and G_r, the total number of potential fragment start positions. G_r is roughly equivalent to twice the number of nucleotides in the genome in that sequences on both the forward and reverse strand are counted. Substituting these definitions into the definition of $Y_{\mathbf{s}}$ gives:

$$Y_{\mathbf{s}} = \frac{p(\mathbf{s}|\theta)}{p(\mathbf{s})} = \frac{C_{\mathbf{s}}}{C_{tot}} \Big/ \frac{N_{\mathbf{s}}}{G_r} \tag{2.6}$$

This provides a third way of thinking of $Y_{\mathbf{s}}$ which is the ratio of $C_{\mathbf{s}}$, the number of fragments associated with a specific sequence, to $C_{tot}N_{\mathbf{s}}/G_r$, the number that would be expected if the fragment starts were uniformly distributed.

$$Y_{\mathbf{s}} = \frac{C_{\mathbf{s}}}{C_{tot}} \Big/ \frac{N_{\mathbf{s}}}{G_r} = C_{\mathbf{s}} \Big/ \frac{C_{tot}N_{\mathbf{s}}}{G_r} \tag{2.7}$$

Figure 2-2 is a toy example showing the calculation of the values of $Y_{\mathbf{s}}$ for a selection of dinucleotides in a 17 nucleotide long genome.



Dinucleotide	*Ctot*	*Gr*	*Ns*	*Cs*	*Ys*
AC	24	16	2	3	1.00
AT	24	16	1	2	1.33
CC	24	16	1	1	0.67
etc	24	16	...	...	...

Figure 2-2 Simple demonstration of the derivation of $Y_{\mathbf{s}}$. The relationship between 2-mer sequences (not 8-mers) and the likelihood of fragmentation is being examined. G_r the number of potential break locations is 16, and C_{tot}, the number of fragments, is 24. Arrows indicate the number of fragments starting at each position. The values of $Y_{\mathbf{s}}$ indicate that the number of breaks associated with the dinucleotide compared to the number that would be expected if the breaks were uniformly distributed.

In the analysed data from ChIP-seq experiments $Y_{\mathbf{s}}$ varies between zero and values significantly greater than one. For example, in the case of the SL523 data, the 8-mer TCGCCGAT occurs 1221 times in the genome, and 145 fragments start in the middle of one

of these 8-mers, where only 3.6 would have been expected given a random distribution, giving a Y_s value of approximately 40.

In this study all alignments exclude any sequences that are deemed unmappable because they align to multiple locations in the genome. The counts N_s and G_r therefore exclude all such unmappable locations. Appendix A provides more details of the algorithm developed to identify these regions.

2.2.3 Log-normal distribution of Y_s

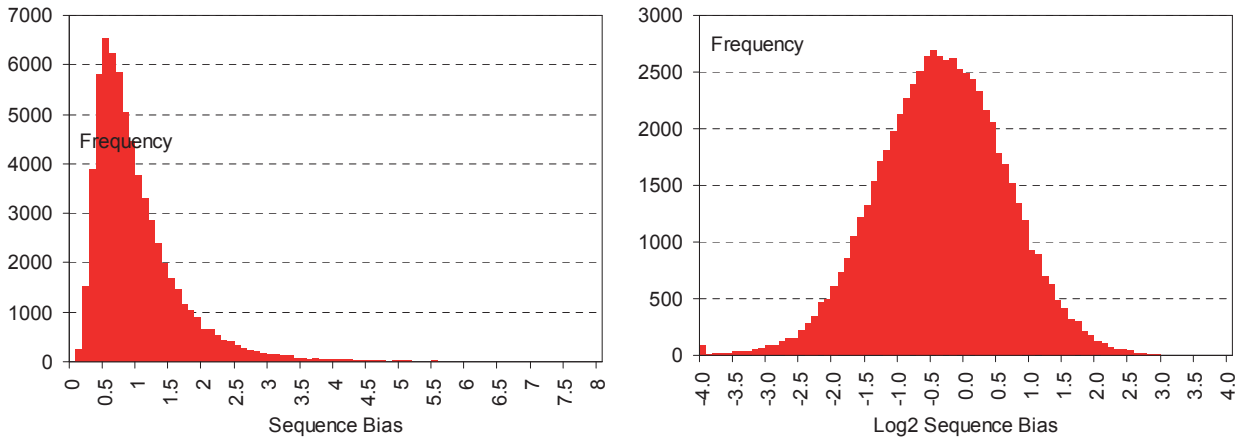


Figure 2-3 Distribution of Ys shows a log normal characteristic. a) Histogram of Y_s for SL523 showing typical skewed distribution. b) Histogram of $\log_2(Y_s)$ showing an approximately normal distribution.

The distribution of Ys has the characteristic of a log-normal distribution (Figure 2-3a), an observation that is supported by the approximately Gaussian distribution of the $\log_2(Y_s)$ values (Figure 2-3b). Consequently base 2 logarithmic scales were used to plot results and $\log_2(Y_s)$ values were used in statistical analyses. If Y_s is viewed as the ratio between the number of fragments associated with a sequence in the experimental data compared to those expected if the distribution was uniform, the use of log values ensures that similar emphasis is given to over- and underrepresented sequences.

2.2.4 Using mutual information to measure the contribution of each nucleotide to the sequence bias ‡

The values of Y_s indicate the influence an eight nucleotide sequence in the genome has on the probability of the DNA fragmenting at a specific position within the sequence. The values of Y_s can then be used to determine the significance of each of the individual nucleotides in determining that probability.

The general principle used to assess the significance of each specific nucleotide is to consider the mutual information of the sequence bias for all the sets of four sequences which only differ in the nucleotide at the position being investigated.

Let S be the 8-mer sequence associated with each specific location in the genome and F the fragment state associated with that position (which can be fragment or not-fragment). If the sequence has no influence on the probability of fragmentation then the two will be independent. The degree of independence can be measured using the mutual information $I(F;S)$ (2.8).

$$I(F;S) = \sum_{f \in F} \sum_{\mathbf{s} \in S} p(f,\mathbf{s}) \log_2 \left(\frac{p(f,\mathbf{s})}{p(f)\,p(\mathbf{s})} \right) \tag{2.8}$$

$p(\mathbf{s})$ is the marginal probability of finding the sequence **s** at any specific location in the genome and $p(f)$ is the marginal probability of a particular fragment state. $p(f,\mathbf{s})$ is the joint probability of a particular fragment state and a specific sequence at any specific location. Complete independence would mean the sequence has no influence on the probability of fragmentation and the mutual information will tend to zero.

Another way of assessing independence is the Kullback-Leibler divergence $D_{KL}\left(p(f|\mathbf{s}) \| p(f)\right)$ between the probability of a specific fragment state given the sequence $p(f|\mathbf{s})$ and the probability of a specific fragment state $p(f)$ which gives a measure of the distance between the two probabilities (2.9). If the probabilities are the same then the distance between the two is zero.

$$D_{KL}\left(p(f|\mathbf{s}) \| p(f)\right) = \sum p(f|\mathbf{s}) \log_2 \left(\frac{p(f|\mathbf{s})}{p(f)} \right) \tag{2.9}$$

Equation (2.8) can be reworked into a form of equation that is very similar to the Kullback-Leibler equation and shows the linkage between the different ways of measuring the degree of dependency of fragmentation on the sequence (2.10).

$$\begin{aligned} I(F;S) &= \sum_{f} \sum_{\mathbf{s}} p(f,\mathbf{s}) \log_2 \left(\frac{p(f,\mathbf{s})}{p(f)\,p(\mathbf{s})} \right) \\ &= \sum_{f} \sum_{\mathbf{s}} p(f|\mathbf{s})\,p(\mathbf{s}) \log_2 \left(\frac{p(f|\mathbf{s})\,p(\mathbf{s})}{p(f)\,p(\mathbf{s})} \right) \\ &= \sum_{f} \sum_{\mathbf{s}} p(f|\mathbf{s})\,p(\mathbf{s}) \log_2 \left(\frac{p(f|\mathbf{s})}{p(f)} \right) \end{aligned} \tag{2.10}$$

The two values directly derivable from the available data are $p(\mathbf{s})$, the probability of a specific sequence and $p(\mathbf{s}|f_1)$ the probability of a finding a specific sequence associated with a fragment. (2.8) can be reworked to use as far as possible the values derivable from the data (2.11)

$$I(F;S) = \sum_{f\in F}\sum_{\mathbf{s}\in S} p(\mathbf{s}|f)\,p(f)\log_2\left(\frac{p(\mathbf{s}|f)\,p(f)}{p(f)\,p(\mathbf{s})}\right)$$
$$= \sum_{f\in F}\sum_{\mathbf{s}\in S} p(\mathbf{s}|f)\,p(f)\log_2\left(\frac{p(\mathbf{s}|f)}{p(\mathbf{s})}\right) \qquad (2.11)$$

The summation over F covers two values, the first, $p(f_1)$, being the probability of finding a fragment, and the second, $p(f_0)$, the probability of there not being a fragment.

The problem with the second value is that it is unknown as the experimental technique does not provide an absolute measure of the number of fragments associated with a specific instance of a region of the genome so it is not possible to quantify $p(f_0)$, the probability of no fragments being associated with a location as all that is known is the relative probabilities of a fragment being associated with a location. Consequently, it is only possible to calculate the mutual information associated with the probabilities of a fragment occurring at a specific location.

However, if it is assumed that the probability of a fragment is small in absolute terms then $p(\mathbf{s}|f_0) \approx p(\mathbf{s})$. The mutual information associated with $p(f_0)$ will therefore tend to zero and the mutual information calculated just using $p(f_1)$ will tend to the value calculated using both probabilities.

If only $p(f_1)$ is being considered then $\sum_f p(f) = p(f_1)$. This value will be a constant that is dependent on the overall number of fragments that were produced in any given experiment. $I'(F;S)$, a normalised variant of the equation will therefore be used where this factor is removed from the equation such that the result is independent of the fragment density (2.12).

$$I'(F;S) = \sum_{\mathbf{s}} p(\mathbf{s}|f_1)\log_2\left(\frac{p(\mathbf{s}|f_1)}{p(\mathbf{s})}\right) \qquad (2.12)$$

In order to assess the significance of the specific nucleotide, the analysis is restricted to the data from genomic locations where one of the set of four sequences $\mathbf{S}_4$ are found. When considering the significance of a specific nucleotide position within the eight nucleotide sequence, there will be 4^{N-1}=16384 sets of four sequences **S** for which the mutual information can be calculated. The set of values can then be used to give an indication as to how much the specific position influences the sequence bias.

Consider the nucleotide sequence **s** consisting of N nucleotides *n*, where $n \in \{\mathrm{a},\mathrm{c},\mathrm{t},\mathrm{g},\phi\}$ and ϕ indicates that no nucleotide value has been defined, such that

$$\mathbf{s} = (n_1\, n_3 \ldots n_N) \tag{2.13}$$

For this analysis of ChIP-seq data N=8. We wish to consider the contribution of n_i, the nucleotide at position *i*. Let $\mathbf{z}(i,j)$ be the j^{th} of the set 4^{N-1} different combinations of the nucleotides in **s** where $n_i = \phi$, $n_x \in \{\mathrm{a},\mathrm{c},\mathrm{t},\mathrm{g}\} | x \neq i$ and $1 \leq j < 4^{N-1}$. The sequence atgϕatgg is an example of $\mathbf{z}(i,j)$ where i = 3. Let $\mathbf{Z}_i$ be the set of all 4^{N-1} of the sequences with a specific value of *i*. This can be represented as:

$$\mathbf{Z}_i = \left\{ \mathbf{z}(i,j) \middle| j \in \{1..4^{N-1}\}, (n_x \in \{\mathrm{a},\mathrm{c},\mathrm{t},\mathrm{g}\} | x \neq i), n_i = \phi \right\} \tag{2.14}$$

For each of the 4^{N-1} sequences in $\mathbf{Z}$ we define $\mathbf{S}_4(i,j)$ the set of four sequences where n_i takes each of the four possible values of *n* for the nucleotide i, (e.g. {"atgaatgg" , "atgcatgg" , "atggatgg" , "atgtatgg"}).

$M(i,j)$ is then defined as the normalised mutual information $I'(F;S)_{i,j}$ for a set of four 8-mers $\mathbf{S}_4(i,j)$:

$$M(i,j) = I'(F;S)_{i,j} = \sum_{\mathbf{s} \in \mathbf{S}_4(i,j)} p(\mathbf{s}|F) \log_2 \left(\frac{p(\mathbf{s}|F)}{p(\mathbf{s})} \right) \tag{2.15}$$

The data from the four sequences aaaaaaaa, caaaaaaa, gaaaaaa and taaaaaa from the SL523 dataset will be taken as a worked example:

Sequence	Seqs	Frags	p(s|F)=a	p(s)=b	a/b=c	log(c)=d	a*c
aaaaaaaa	6666259	20000	0.6349	0.7080	0.8968	-0.1572	-0.0998
caaaaaaa	830530	2836	0.0900	0.0882	1.0206	0.0295	0.0027
gaaaaaaa	994995	5644	0.1792	0.1057	1.6955	0.7617	0.1365
taaaaaaa	923224	3019	0.0958	0.0981	0.9774	-0.0330	-0.0032
	9415008	**31499**					0.0361
						M(*i*,*j*) =	0.0361

This gives the mutual information for one of the sets of four sequences associated with the nucleotide at position *i*. The operation is repeated for all of the other sets of four sequences.

The distribution of the mutual information values derived in this way for a given nucleotide shows a skewed distribution that is similar to a log-normal distribution (Figure 2-4a) suggesting that it might be appropriate to work with the logarithms of these values when looking at variations and distributions (Figure 2-4b).

The average mutual information $\bar{R}(i)$ was calculated using the log values and then converted back to an absolute value, which is equivalent to the geometric mean of the absolute values (2.16).

$$\bar{R}(i) = \exp\left(\frac{\sum_{j=1}^{N} \log\left(M(i,j)\right)}{N}\right) = \sqrt[N]{\prod_{j=1}^{N} M(i,j)} \tag{2.16}$$

The value of $\bar{R}(i)$ indicates the degree of interaction between the probability of a DNA fragment starting at a particular position and the nucleotide type at position *i* relative to the fragment start. The average value is numerically small so the interaction intensity (*II*) has been defined where

$$II(i) = \bar{R}(i) \times 10^6 \tag{2.17}$$

A larger value of *II*(*i*) indicates a greater dependence of the fragmentation probability on the nucleotide at position *i*.

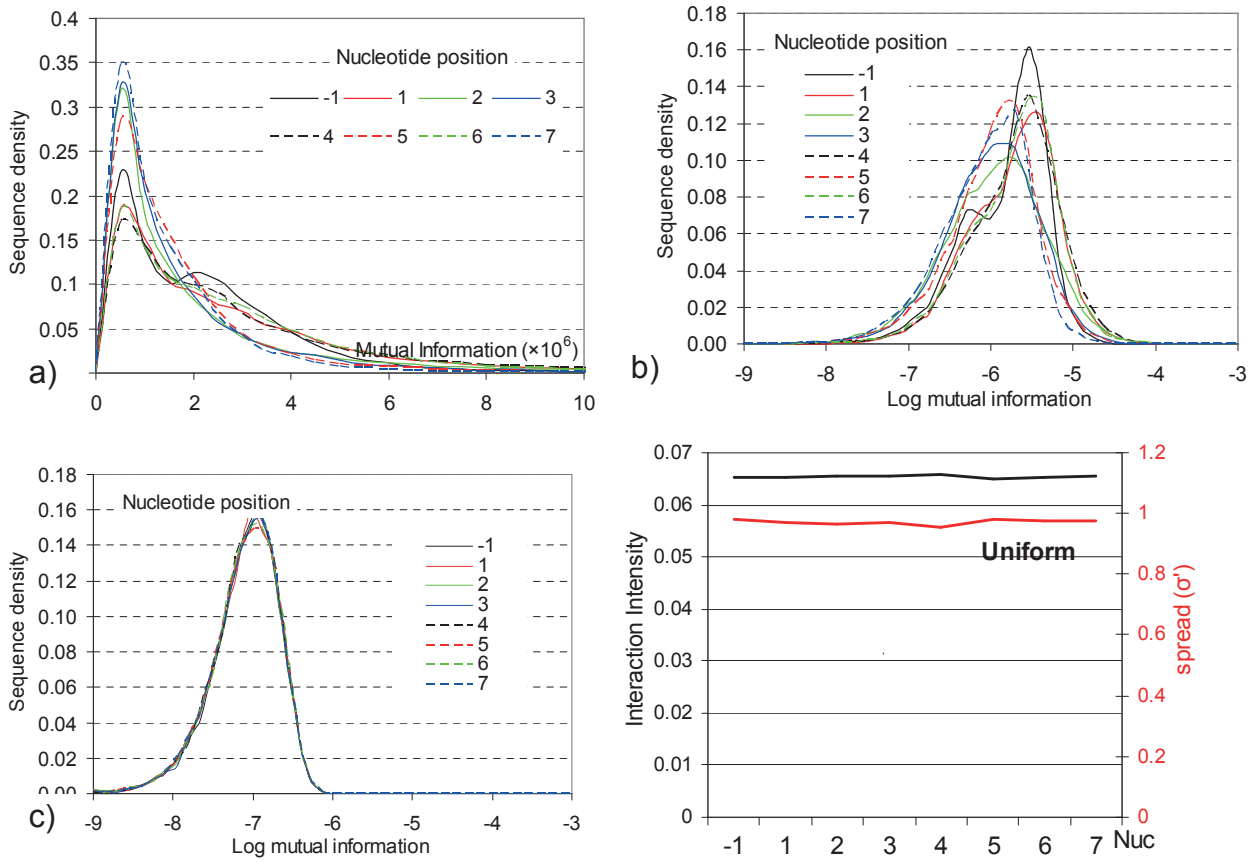


Figure 2-4 Distribution of mutual information for the sequence sets associated with each nucleotide shows a log normal characteristic. Both graphs show the distribution of the mutual information for each of the 16384 sets of four 8-mers associated with each nucleotide in the SL117 dataset a) Distribution of values showing skewed distribution b) Distribution of log values showing a greater tendency to a Gaussian distribution. c) Distribution for uniform dataset, showing significantly lower average values than the SL117 data. d) Interaction Intensity and spread for uniform dataset

While the distribution of the mutual information for some nucleotides shows some interesting bimodal characteristics, it was decided just to use a simple measure of the spread of the values to identify additional nucleotide by nucleotide variation in the mutual information. This was obtained by calculating the multiplicative standard deviation or standard deviation of the log values $\sigma'\left(R(i)\right)$ (2.18).

$$\sigma'\left(R(i)\right) = \sqrt{\frac{\sum_{j=1}^{N}\left(\log\left(M(i,j)\right) - \log\left(\bar{R}(i)\right)\right)^2}{N-1}} = \sqrt{\frac{\sum_{j=1}^{N}\left(\log\left(M(i,j)\,/\,\bar{R}(i)\right)\right)^2}{N-1}} \qquad (2.18)$$

This value is an indication of how much the interaction intensity varies depending on the neighbouring nucleotide sequence. A lower value indicates that there is less variation, i.e. the contribution is largely independent of the neighbouring nucleotides. A larger value indicates a greater interdependence between the nucleotide at this position and the neighbouring sequence when determining the probability of a fragment start.

The calculation of mutual information, and therefore the interaction intensity and the spread makes no assumptions about any specific relationship between nucleotide bias and sequence. For example, it could be that n_i=T sometimes increases the probability of DNA fragmentation and sometimes n_i=A increases the probability, this being dependent on the adjacent nucleotides. Both cases will however contribute to the Interaction Intensity.

The values for a uniform dataset (figure 2-4c and d) gives a baseline which can be used when assessing the values of experimental dataset.

2.2.5 Representation of bias using Position Coefficient Matrixes PCMs ‡

The convention adopted throughout this research is to represent the bias of a set of nucleotides within a sequence of length N as a Position Coefficient Matrix (PCM) **M** such that

$$\begin{aligned} \mathbf{M} &= \left(\mathbf{m}_1, \mathbf{m}_2, \mathbf{m}_3 .. \mathbf{m}_N\right) \\ \mathbf{m}_i &= \left(\chi_{i,\mathrm{a}}, \chi_{i,\mathrm{c}}, \chi_{i,\mathrm{g}}, \chi_{i,\mathrm{t}}\right) \end{aligned} \tag{2.19}$$

$\chi_{i,n}$ is the coefficient associated with the nucleotide of type n at position i, defined such that

$$\sum_{n\in\{\mathrm{a,c,g,t}\}} \chi_{i,n} = 1 \tag{2.20}$$

If there is no bias associated with the nucleotide at position i then the coefficients will all equal 0.25.

This is similar to the use of Position Specific Scoring Matrices (PSSMs) to describe the frequency with which each nucleotide is found at each position in a genomic motif such as a transcription factor binding site (Section 1.4.11) [51]. The similarities between these two conventions allows the use of standard techniques such using logos to display the values in a PCM [87].

It should be stressed however that in the model used in thesis, unlike when PSSMs are used to describe transcription binding sites, $\chi_{i,n}$ does not represent the relative frequency with which a nucleotide is found at a location. It is instead a measure of the sequence bias at the

location, which is a measure of how much different nucleotides at that location affect the likelihood of a fragment start occurring nearby. This model will be described in more detail in Section 2.2.7. It is nevertheless the case that larger values of $\chi_{i,n}$ indicate that there is a greater probability of nucleotide *n* occurring at position *i* in the sequence described by the PCM.

Model fitting is used to determine the values of **M** that give the closest match between the observed data and the model.

2.2.6 Mapping nucleotide weights to three dimensional vectors ‡

The convention that the weights at a specific location sum to unity introduces a redundancy into the four PCM weights $\chi_{i,n}$, in that knowing the values of three of the weights automatically determines the fourth weight. Model fitting is more efficient if such redundancy is removed, such that the model fitting algorithm is working with the smallest possible set of model fitting variables. This could be done by using the model fitting algorithm to vary three of the four weights and deriving the fourth from the other three. However, the approach that was adopted was to map the four weights at each position in the PCM to a three dimensional (3D) vector as follows:

$$\begin{aligned}\tilde{\chi}_{\mathrm{x}} &= \frac{\left(\chi_{\mathrm{c}}+\chi_{\mathrm{g}}\right)-\left(\chi_{\mathrm{a}}+\chi_{\mathrm{t}}\right)}{\chi_{\mathrm{a}}+\chi_{\mathrm{c}}+\chi_{\mathrm{g}}+\chi_{\mathrm{t}}} \\ &= \left(\chi_{\mathrm{c}}+\chi_{\mathrm{g}}\right)-\left(\chi_{\mathrm{a}}+\chi_{\mathrm{t}}\right) \qquad \tilde{\chi}_{\mathrm{y}} = \frac{\chi_{\mathrm{a}}-\chi_{\mathrm{t}}}{\chi_{\mathrm{a}}+\chi_{\mathrm{t}}} \qquad \tilde{\chi}_{\mathrm{z}} = \frac{\chi_{\mathrm{c}}-\chi_{\mathrm{g}}}{\chi_{\mathrm{c}}+\chi_{\mathrm{g}}}\end{aligned} \tag{2.21}$$

$$\tilde{\mathbf{m}} = \left(\tilde{\chi}_{\mathrm{x}}, \tilde{\chi}_{\mathrm{y}}, \tilde{\chi}_{\mathrm{z}}\right) \tag{2.22}$$

This can be visualised as mapping the coefficients **m** onto a vector $\tilde{\mathbf{m}}$ in 3D space starting at the origin. A feature of this mapping is that all valid combinations of nucleotide weights $\sum \chi_n = 1,\ 0 < \chi_n < 1$ map to a vector ending within a cube of edge length two which is centred on the origin. This mapping treats each nucleotide type in an equivalent way. The reverse mapping is as follows:

$$\begin{aligned}\chi_{\mathrm{a}} &= \tfrac{1}{4}\left(1-\tilde{\chi}_{\mathrm{x}}\right)\left(1+\tilde{\chi}_{\mathrm{y}}\right) & \chi_{\mathrm{c}} &= \tfrac{1}{4}\left(1+\tilde{\chi}_{\mathrm{x}}\right)\left(1+\tilde{\chi}_{\mathrm{z}}\right) \\ \chi_{\mathrm{g}} &= \tfrac{1}{4}\left(1+\tilde{\chi}_{\mathrm{x}}\right)\left(1-\tilde{\chi}_{\mathrm{z}}\right) & z_{\mathrm{t}} &= \tfrac{1}{4}\left(1-\tilde{\chi}_{\mathrm{x}}\right)\left(1-\tilde{\chi}_{\mathrm{y}}\right)\end{aligned} \tag{2.23}$$

From this definition of a 3D representation of the weights at any specific position, $\tilde{\mathbf{M}}$, an ordered set of such vectors covering all of the nucleotides in the PCM can then be defined as follows:

$$\tilde{\mathbf{M}} = \left(\tilde{\mathbf{m}}_1, \tilde{\mathbf{m}}_2, \tilde{\mathbf{m}}_3 .. \tilde{\mathbf{m}}_N\right), \text{ where } \tilde{\mathbf{m}}_i = \left(\tilde{\chi}_{i,\mathrm{x}}, \tilde{\chi}_{i,\mathrm{y}}, \tilde{\chi}_{i,\mathrm{z}}\right) \tag{2.24}$$

A location within the PCM where the nucleotide type has no effect on the characteristic being modelled $\left(\chi_{i,n} = 0.25 \ \forall n\right)$ maps to a vector of length zero $\left(\tilde{\chi}_{i,m} = 0 \ \forall m\right)$. The length of the vector is an indication of the extent of the influence of the nucleotide on fragmentation.

When the redundancy was removed in this way by mapping to a 3D vector, model fitting was found to proceed at a very similar rate to simpler mapping where only the values for three nucleotide are optimised and the fourth is simply derived (Data not shown). The results of model fitting were also shown to be mapping independent (Data not shown).

As well as being used to reduce the model fitting problem to the minimum number of independent coefficients, the mapping is also used to map the PCM to a vector space where vector operations such as scaling and the addition of vectors associated with a set of matrices can be performed.

2.2.7 Modelling sequence bias using one or more PCMs †

Position Coefficient Matrices (PCMs) are used to model the pattern of nucleotide bias within the N-mer, where the matrix covering N nucleotides can then be used to calculate a predicted or modelled bias $M_{\mathbf{s}}$ given by:

$$M_{\mathbf{s}} = k \prod_{i=1}^{N} 4\chi_{i,n_i} \tag{2.25}$$

This is analogous to the use of PSSM weights to calculate the likelihood of a specific sequence. χ_{i,n_i} is the positive coefficient associated with the nucleotide n_i at position i and k a scalar multiplier.

The factor of four ensures that k and the coefficients are independent of the number of coefficients, e.g. adding an additional null coefficient at position j (i.e. $\chi_{j,n} = 0.25 \forall n$) does not change the value of $M_{\mathbf{s}}$. The rationale for this model is the hypothesis that nucleotide at each position makes an independent contribution to the likelihood of DNA fragmentation.

In order to model bias as multiple alternative sequence patterns, the model was extended to incorporate the possibility of P alternative PCMs each with a scalar multiplier k_j, plus a single global offset parameter o such that

$$M_{\mathbf{s}} = \max\left(k_j \prod_{i=1}^{N} 4\chi_{i,n_i,j}\right)_{j=1}^{P} + o \tag{2.26}$$

$M_{\mathbf{s}}$ is used as a predictor of $Y_{\mathbf{s}}$ and the values of $\chi_{i,n_i,j}$ and k_j were optimised in order to achieve the best match between $Y_{\mathbf{s}}$ and $M_{\mathbf{s}}$. The distance between the logarithms of $Y_{\mathbf{s}}$ and $M_{\mathbf{s}}$

was minimised in view of the log-normal distribution of $Y_{\mathbf{s}}$ (Section 2.2.3) in order to give a balanced weighting to the effect of sequences that are both under and over-represented. The error function E_{DNA} used for model optimisation is the sum of the squares of the difference between the logarithms of $Y_{\mathbf{s}}$ and $M_{\mathbf{s}}$ (2.27).

$$\begin{aligned} E_{DNA} &= \sum_{\mathbf{s}:N_{\mathbf{s}}>T} \left(\log_2\left(M_{\mathbf{s}}\right) - \log_2\left(Y_{\mathbf{s}}\right)\right)^2 \\ &= \sum_{\mathbf{s}:N_{\mathbf{s}}>T} \left(\log_2\left(M_{\mathbf{s}} \middle/ Y_{\mathbf{s}}\right)\right)^2 \end{aligned} \tag{2.27}$$

T is a threshold which removes the contribution from N-mers where there are only a small number of instances in the genome. This reduces the noise contribution from data associated with these N-mers and also the computational load during model fitting. In this analysis using 8-mers and the human genome, T was set to 50000 which retains 55% of the sequences, and only excludes contributions from 0.03% of the genome because of the sparsity of these 8-mers. It was confirmed that the model fitting was robust under variation of T.

The rational for this extension to the model is the hypothesis that there are alternative independent sequence patterns that are associated with an increased likelihood of DNA fracture, and the likelihood of DNA fracture at a particular location is determined by the pattern that most closely matches the DNA sequence at that position.

After optimisation, an x-y plot of the log values is used to show the closeness of fit of the model and the observed data.

2.2.8 Model fitting using the Nelder-Mead function minimisation algorithm ‡

The model being considered includes the effect of DNA sequences of up to eight nucleotides in length on the likelihood of DNA fragmentation, and each nucleotide has three associated parameters making a total of 24 parameters per sequence pattern, represented by a PCM. Up to eight independent alternative PCMs, each with an associated scalar are incorporated into the model, generating a total of 200 model fitting parameters.

As well as having of the order of 200 independent parameters, the function being fitted is very non-linear, in that the function for any given DNA sequence is the maximum of a set of possible values associated with the set of PCMs. A small change in a parameter can result in a change in the identity of the PCM that defines the function result for a significant number of DNA sequences, creating non-linearities in the derivatives of the error function. Such non-linearities can cause problems for some model fitting algorithms.

The model fitting algorithm used was an extension of the Amoeba optimisation algorithm [34], based on the Nelder-Mead function minimisation algorithm [74]. This algorithm is well suited to problems where there are large numbers of parameters to be optimised, and where there are significant non-linearities in the function being optimised in that it does not rely on the existence of well behaved differentials.

The coefficients of the model being fitted can be represented by a vector **c** where

$$\mathbf{c} = \left(k_1, k_2..k_m, \tilde{\mathbf{M}}_1, \tilde{\mathbf{M}}_2..\tilde{\mathbf{M}}_m\right) = \left(c_1, c_2..c_q\right) \tag{2.28}$$

The function minimisation is done by creating an initial vector $\mathbf{c}_0$ where all of the parameters set to an appropriate initial value, and also a set of q vectors $\mathbf{c}_n$.

$$\begin{aligned}
\mathbf{c}_o &= \left(c_1, c_2, c_3, \ldots, c_q\right) \\
\mathbf{c}_1 &= \left(c_1 + \delta_1, c_1, c_2, \ldots, c_q\right) \\
\mathbf{c}_n &= \left(c_0, c_1, \ldots, c_{n-1}, c_n + \delta_n, c_{n+1}, \ldots, c_q\right) \\
\mathbf{c}_q &= \left(c_0, c_1, \ldots, c_q + \delta_q\right)
\end{aligned} \tag{2.29}$$

These vectors are such that the set $\left\{\mathbf{c}_1 - \mathbf{c}_0, \mathbf{c}_2 - \mathbf{c}_0, \mathbf{c}_3 - \mathbf{c}_0, \ldots, \mathbf{c}_n - \mathbf{c}_0, \ldots, \mathbf{c}_q - \mathbf{c}_0\right\}$ are orthogonal.

Model fitting proceeds by minimising an error value E which is a measure of the mismatch between the experimental results and model predictions. In the case of the ChIP-seq data the model fitting algorithm searches for the PCM coefficients that minimise E_{DNA} so $E = E_{\mathrm{DNA}}$.

The error value E is then calculated for each vector, and these are used to obtain an indication of the local characteristics of E in the multidimensional vector space, which informs the creation of a new vector $\mathbf{c}_\mathrm{v}$ which is considered as a possible replacement for the vector $\mathbf{c}_\mathrm{h}$ which gave the highest value of E_r where

$$\mathbf{c}_r = \overline{\mathbf{c}} + \left(\overline{\mathbf{c}} - \mathbf{c}_h\right) \tag{2.30}$$

$\overline{\mathbf{c}}$ is the average of the q+1 vectors $\mathbf{c}_\mathrm{n}$. The associated sum of squares error E_r is calculated, and compared with the lowest error E_l. If $E_v < E_l$, then two further vectors are generated

$$\mathbf{c}'_r = \overline{\mathbf{c}} + 2\left(\overline{\mathbf{c}} - \mathbf{c}_h\right), \qquad \mathbf{c}''_r = \mathrm{RC}\left(\mathbf{c}_r, \mathbf{c}'_r\right) \tag{2.31}$$

For each of these the errors E'_v and E''_v are derived. The operation RC(**x**,**y**) creates a new vector where each elements is selected randomly from co-positional elements of the two vectors **x** and **y**. The vector $\mathbf{c}_h$ is then replaced by the best of these vectors, as determined by

the values of E_v, E'_v and E''_v for each of the vectors and the process repeats. The use of the RC function is the extension of the original Amoeba algorithm that was introduced during this research. It introduces a small amount of scaled noise which increases the speed with which the algorithm is able to negotiate significant changes of slope in the multidimensional vector space within which the model fitting takes place.

If E_r was not lower than E_l then a vector $\mathbf{c}'_h$ is created where

$$\mathbf{c}'_h = \bar{\mathbf{c}} + 0.5\left(\mathbf{c}_h - \bar{\mathbf{c}}\right) \tag{2.32}$$

And this is used as a replacement for $\mathbf{c}_h$ providing that it has a lower error value. If this is not the case then a new set of orthogonal vectors are generated centred on the current $\bar{\mathbf{c}}$ This process is represented in Figure 2-5.

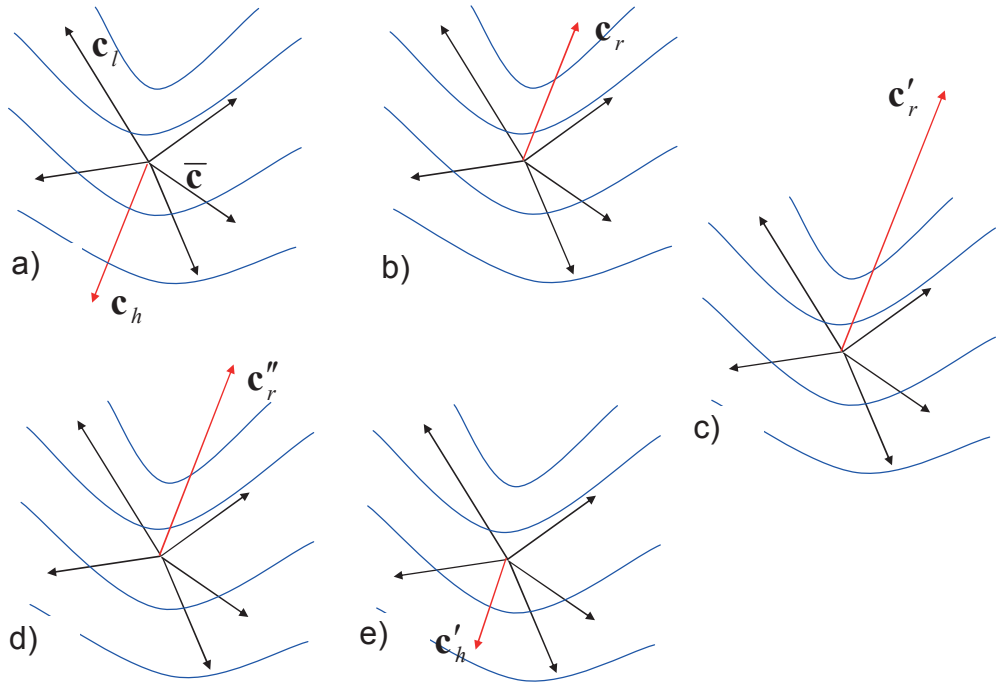


Figure 2-5 Representation of core Nelder-Mead optimisation step a) Set of vectors each representing the difference between a set of parameters and the average $\bar{\mathbf{c}}$ where $\mathbf{c}_h$ is the vector with the highest error. $\mathbf{c}_h$ is replaced by $\mathbf{c}_r$ (b) which is $\mathbf{c}_h$ reflected through $\bar{\mathbf{c}}$ or $\mathbf{c}_r'$ (c) which is twice the length of $\mathbf{c}_r$ or $\mathbf{c}_r''$ (d) which is a randomised mix of $\mathbf{c}_r$ and $\mathbf{c}_r'$ depending on which has the lowest error, providing that it is lower than the error associated with $\mathbf{c}_l$. Otherwise a vector $\mathbf{c}'_h$ which is half of $\mathbf{c}_h$ (e) is used to replace $\mathbf{c}_h$ if it gives a better error than $\mathbf{c}_l$.

The model fitting proceeds with the fitting of a single PCM to the data until a minimum E_{DNA} is achieved. The model is then extended by adding a second PCM with identical parameters to the first optimised PCM. Further model fitting causes the PCMs to diverge from

each other such that one of the PCMs is selected by the algorithm for some of the sequences and the second for the other, depending on the associated value of M_s. The divergence is driven by the fact that a better model fit can be achieved for a set of sequences that are associated with a single PCM if two PCMs are used instead. The single PCM can be considered as taking values that are a compromise between the values of the two PCMs. Once an optimal fit has been reached with two PCMs, the results are examined to see which PCM has been adopted for the majority of the sequences, and a copy of this PCM is then added to the model, and the process repeated. This is a form of clustering where the n^{th} cluster consists of the set of sequences such that the value of $k_j \prod_{i=1}^{N} 4\chi_{i,n_i,j}$ is largest when $j = n$ (Equation (2.26)). Additional clusters are added by taking the cluster with the greatest number of sequences and attempting to split it by adding an additional PCM.

The decision as to when the model fitting has reached a minimum E_{DNA} at each stage was done by examining the record of the improvement in the E_{DNA} to determine when a plateau had been reached.

2.2.9 Zooming into PCMs in order to make the bias visible in logos ‡

There were a number of cases in this investigation where the bias of each nucleotide appeared to be relatively small such that the values of $\chi_{i,n}$ remain relative close to 0.25, and the equivalent 3D vector is relatively small. When this happens, the representation of the values as a logo using the previously published algorithm [87] is very uninformative as the overall character height at a nucleotide position is determined by the information content of the set of coefficients for the nucleotide. The effect of this small bias is nevertheless significant when it is considered in combination with the bias of all the other nucleotides.

A technique was introduced to magnify the information content of the PCM by mapping the weights to their 3D vector representation and scaling the lengths of all the vectors in the PCM (or set of PCMs) by the same factor and then mapping the vector back to the nucleotide weights. The magnification factor or ‘Zoom’ that was chosen was such that the information contained within the PCM is clearly visible in the logo. The ‘Zoom’ factor applied is recorded on the logo. Figure 2-6 demonstrates the effect of applying zoom to a PCM.

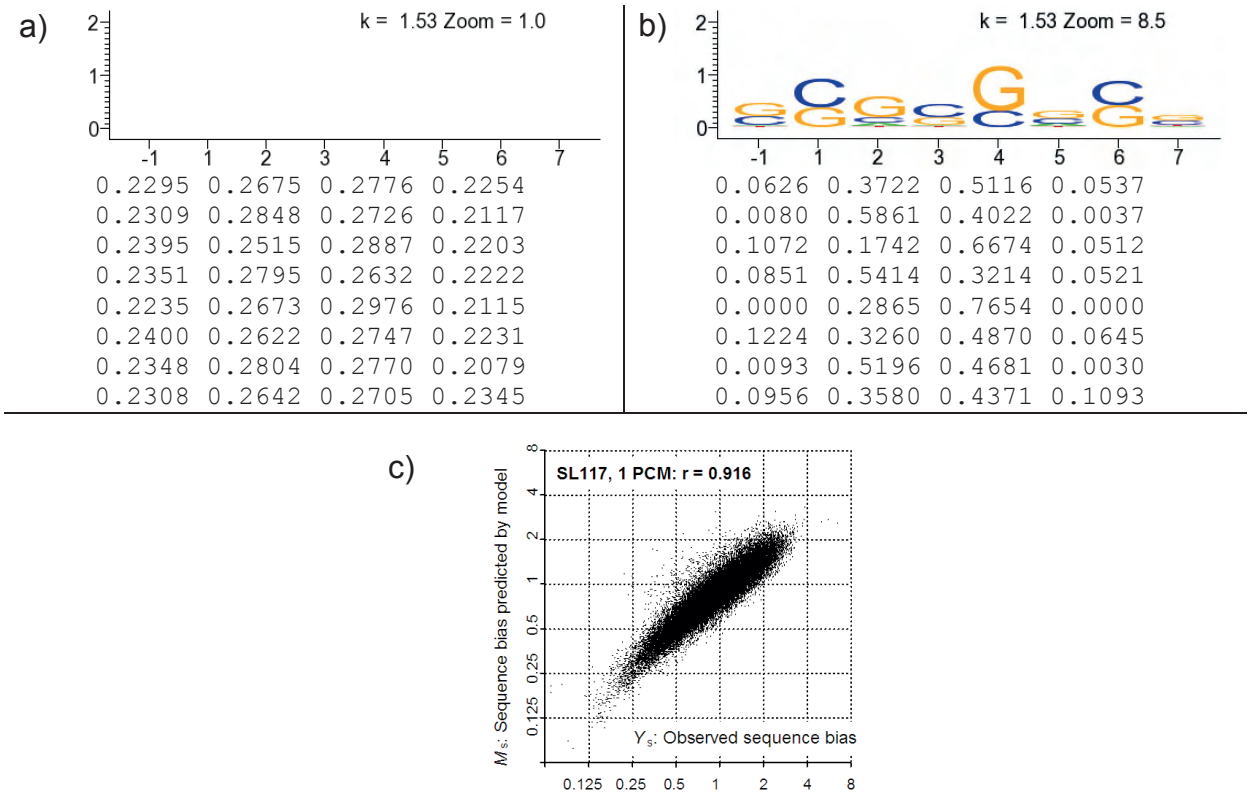


Figure 2-6 An example of the use of 'Zoom' when displaying logos. a) The PCM created by model-fitting that results in the fit between observed data and model shown in c). The values are sufficiently close to 0.25 that the information content scaling factor results in a non-informative logo. b) shows the data and the associated logo after a zoom of 8.5 has been applied to the weights in order to create a logo that highlights the information in the original PCM. c) Plots the $Y_{\mathbf{s}}$ and $M_{\mathbf{s}}$ for each of the 8-mer sequences to demonstrate that the cumulative effect of the underlying biases is significant enough to give a good model fit.

2.2.10 Adjusting for sequence bias

The availability of sequence bias information for a set of data creates the possibility of adjusting the data to compensate for the sequence bias and so reducing the degree to which it might mask other factors that affect the fragment distribution.

Let the N-mer associated with the position x in the genome be $\mathbf{s}_x$ and f_x be the number of fragment starts at x. The general approach adopted to adjust for sequence bias is to multiply the number of fragment starts by a factor that adjusts for the degree that fragments associated with the N-mer are under- or overrepresented.

Two ways for adjusting f_x have been used with the data examined in this thesis. The first was to use $Y_{\mathbf{s}(x)}$, the sequence bias for the N-mer at location x that has been derived for this dataset as defined in Section 2.1.1. The second was to use the predicted sequence bias $M_{\mathbf{s}}$ for the N-mer at location x that comes from the model created using model fitting as defined in

Section 2.2.7. It would be expected that for most sequences these two values will be similar as it is the objective of model fitting to minimise the difference between these values. These two alternatives are designated f'_x and f''_x and are defined as follows:

$$f'_x = f_x \Big/ Y_{\mathbf{s}(x)} \tag{2.33}$$

$$f''_x = f_x \Big/ M_{\mathbf{s}(x)} \tag{2.34}$$

If there are no breaks associated with an N-mer in the reference dataset then $Y_{\mathbf{s}(x)} = 0$ resulting in an adjusted count of ∞. In such situations, no adjustment is made. This is a very rare occurrence as this only happens for N-mers that rarely have an associated fragment start.

2.3 Primary results and discussion

This section shows the results that are central for the overall conclusions of this investigation. Section 2.4 provides additional results that examine some specific aspects in more detail.

2.3.1 Distribution of ChIP-seq fragment ends

Input data from ChIP-seq experiments were chosen because they will not suffer from any biasing effects that might be associated with the immunoprecipitation stage that comes later in the ChIP-seq protocol.

Figure 2-7 shows a representative distribution of fragment starts and ends from a 10000 nucleotide length region of SL523 and SL117, the two primary sets of ChIP-seq input data used in this investigation. In both cases the fragment locations are relatively evenly distributed through the region shown, qualitatively justifying their use in this investigation. There is however still some evidence of a subtle difference in the character of the fragment distribution, with the SL523 data showing more instances where there is more than one sequence tag associated with a specific location in the genome. As expected, there are fewer fragments in the regions with a greater proportion of non-unique sequences. Non-unique sequences means that it is not possible to map uniquely the sequences to the genome where the convention used in this thesis is that all such fragments will therefore be discarded.

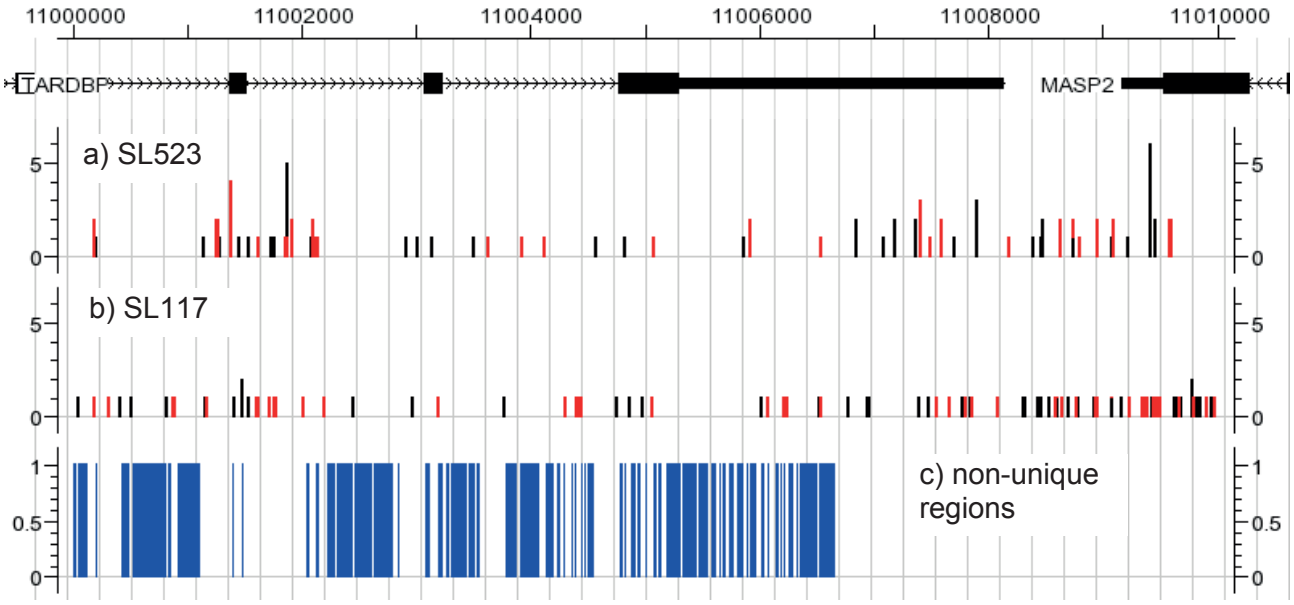


Figure 2-7 Fragment ends for two primary datasets have similar distributions. Distribution of fragment ends for the two primary ChIP-seq datasets used in this investigation for nucleotides 11000000 to 11010000 of chromosome 1. Black bars in a) and b) indicate the number of sequences that aligned to the forward strand, indicating the fragment start position w.r.t. forward strand. Red bars indicate the number of fragments that end at a specific location, Track c) indicates those regions where a 25 base-pair sequence is non-unique so that it is not possible to uniquely map a sequence to the genome, black for forward strand mappings and red for reverse strand mappings.

2.3.2 There is a significant sequence-dependent bias in fragment start locations

The initial investigation looked at the 8-mers centred on the start position of the sequenced fragments and determined the degree to which each of the 65536 8-mers were under- or overrepresented compared to the number that would be expected if the probability of fragmentation was independent of the sequence. A dataset was created to simulate such sequence independence by assuming the fragment starts were spaced at equal distances, or uniformly distributed, throughout the genome.

If $E_{\mathbf{s}}$ is the expected number of fragments associated with a specific sequence given a uniform distribution, then the actual counts in multiple experiments would be expected to have a Poisson distribution where $\lambda = E_{\mathbf{s}}$ and so would have a mean of $E_{\mathbf{s}}$ and a standard deviation of $\sqrt{E_s}$. $Y_{\mathbf{s}}$, which equals $C_{\mathbf{s}}/E_{\mathbf{s}}$, would therefore have a mean of one and a standard deviation of $1/\sqrt{E_{\mathbf{s}}}$. For large values of $E_{\mathbf{s}}$, the distribution tends to a Gaussian distribution. While the average value of $E_{\mathbf{s}}$ is of the order of 200, when the distribution will be very close to Gaussian, there is a considerable variation with the lowest value being of $E_{\mathbf{s}}$ being of the

order of 2. It was decided to normalise each $Y_{\mathbf{s}}$ *to* $Y'_{\mathbf{s}}$ with a mean of zero and a standard deviation of one on the assumption that $Y_{\mathbf{s}}$ has a standard deviation of $1/\sqrt{E_{\mathbf{s}}}$ in order to be able to compare the values for different sequences and assess the degree to which the values were different from those expected with the null, uniform distribution. The mapping is:

$$Y'_{\mathbf{s}} = \left(Y_{\mathbf{s}} - 1\right)\sqrt{E_{\mathbf{s}}} \tag{2.35}$$

SL523

Seq	N_s	C_s	E_s	Y_s	Y'_s	p value
GGGGGGGG	59880	2793	177.6	15.7	196.24	1E-8360
TGTGTGTG	1055108	13520	3129.8	4.3	185.72	1E-7488
TTTGTTTT	932731	11906	2766.8	4.3	173.75	1E-6553
ACTGTACA	89261	2772	264.8	10.5	154.08	1E-5153
CCGATACG	958	238	2.8	83.8	139.50	1E-4223
TTTGTTTG	395503	5842	1173.2	5.0	136.31	1E-4032
TTCTCTGC	435334	315	1291.4	0.2	-27.17	1E-157
ACATCCTT	525802	444	1559.7	0.3	-28.25	1E-170
TGGCCCAG	510305	407	1513.7	0.3	-28.45	1E-173
ATATCAGA	675513	683	2003.8	0.3	-29.51	1E-186
TGTCCTCA	474170	248	1406.6	0.2	-30.89	1E-204
GGAGCAGG	947152	1035	2809.6	0.4	-33.48	1E-241

Uniform

Seq	N_s	C_s	E_s	Y_s	Y'_s	p-value
AGTATGCA	74748	372	299.0	1.2	4.22	2.05E-05
TCCTGGTC	90356	440	361.4	1.2	4.13	2.81E-05
GTAACCGT	6598	47	26.4	1.8	4.01	1.00E-04
TAGCGTGT	10162	66	40.6	1.6	3.98	9.42E-05
TTCAGGAG	188728	864	754.9	1.1	3.97	4.74E-05
TATCGCAA	21210	121	84.8	1.4	3.93	8.76E-05
AAAGAATC	8505	14	34.0	0.4	-3.43	8.84E-05
TAGCGCTG	8558	14	34.2	0.4	-3.46	7.77E-05
TAGCCTTT	158581	547	634.3	0.9	-3.47	2.13E-04
TTTCGAGT	19722	48	78.9	0.6	-3.48	1.23E-04
TCCCTTTA	242755	857	971.0	0.9	-3.66	1.03E-04
ACCACAAG	98917	321	395.7	0.8	-3.75	5.99E-05

Table 2-1 Over- and underrepresented sequences show significant sequence bias. a) List of six most over- and underrepresented sequences surrounding fragment start positions in SL523 showing, for each 8-mer **s**, the number of 8-mers in the genome ($N_{\mathbf{s}}$), the number of breaks associated with the 8-mers ($C_{\mathbf{s}}$), the number that would be expected if the break locations were uniformly distributed ($E_{\mathbf{s}}$) the sequence bias ($Y_{\mathbf{s}}$) and the sequence bias after mapping to a distribution where the null hypothesis gives a mean of zero and a standard deviation of 1.0 ($Y'_{\mathbf{s}}$). The p-value gives the probability of a value of *Ys* that is at least as extreme. b) The equivalent data for an artificial dataset with uniform distribution. The p-values for the uniform distribution are consistent with the null hypothesis in that the values are consistent with the values that would be found at the tails of a distribution with 65536 different sequences. The p-values for the SL523 dataset indicate that the values seen are extremely unlikely given the null hypothesis.

Table 2-1 shows the derivation of the sequence bias for the six sequences which are furthest away on either side of the expected value based on the normalised Y'_s for both the SL523 data and also for a dataset created to show the distribution if the fragments are uniformly distributed in the genome with the same average density as the SL523 data. This shows that in SL523 there is a very wide range of normalised sequence bias values, with some very under- and some very overrepresented sequences. In order to determine whether this is significant or not, cumulative distribution functions can be used to calculate the probability of finding a value at least as extreme to give a p-value.

In the case of the uniform distribution, the values of Y_s are sufficiently small that cumulative distribution can be calculated using the Poisson distribution. The p-values shown are for the six most extreme values in the two tails of the distributions, and the values are consistent with those that would be expected at the tails, in that the values are of the order of 1/60000, and some p-values of this order would be expected in a distribution with 65536 values.

In the case of the SL523 dataset, the values are sufficiently extreme that normal arithmetic precision limitations result in p-values of zero. In most cases the values are sufficient to justify assuming a Gaussian approximation and the cumulative distribution approximation for large values can be used (Appendix F-2). The p-values show that the sequence bias observed is significantly different from that which would be expected given the null distribution.

The distribution of the normalised sequence bias for the artificial uniform dataset was calculated and compared with a normal distribution with the same mean and standard deviation (Figure 2-8). The match is very good, increasing the confidence in the mapping technique used and the assumptions that distribution after such a mapping will be Gaussian.

In the SL523 dataset after normalisation, some sequence biases are 50 standard deviations higher than the expected mean given the null hypothesis, and others are 30 times lower. Most of the biases are significantly outside the range that would be expected of +/- 4 s.d. This shows that the distribution of fragment starts within the 8-mers is significantly different from what would be expected if the fragment distribution was sequence independent, as simulated by the uniform fragment distribution.

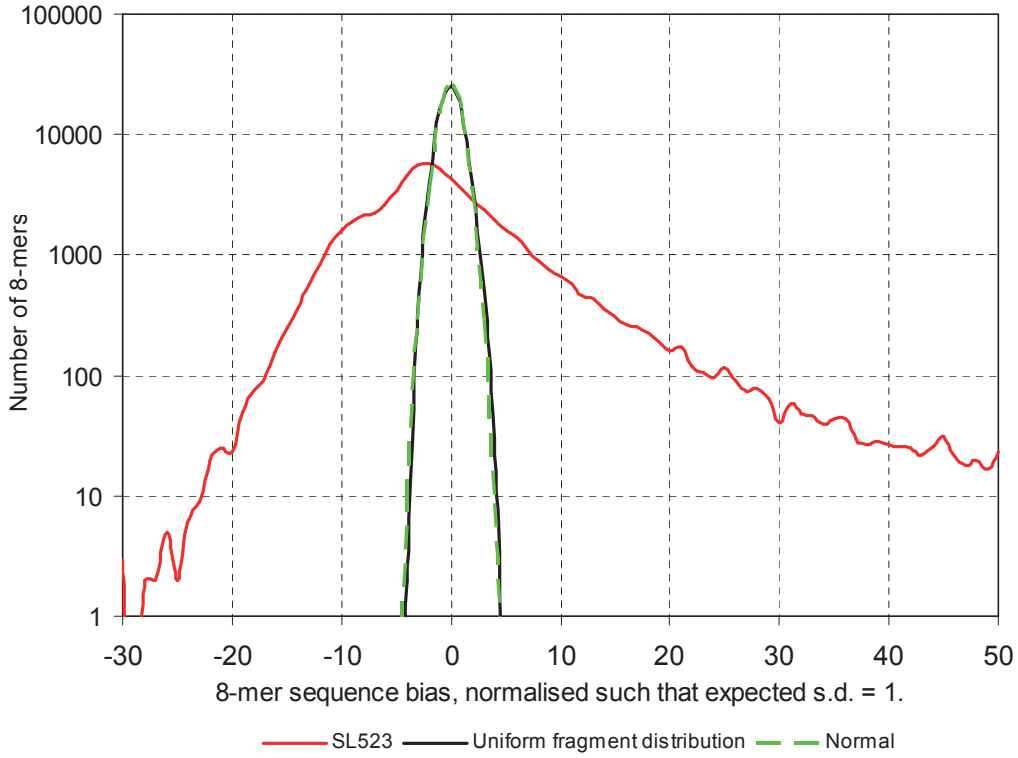


Figure 2-8 Sequence bias distribution for SL523 is significantly different from that of a uniform fragment distribution. The normalised SL523 bias distribution (red) is significantly different from an artificial dataset with a uniform fragment distribution after normalisation (black), which is very close to the expected normal distribution (black).The SL523 distribution shows biases that are 50 standard deviations higher and 20 standard deviations lower than would be expected assuming a uniform distribution.

2.3.3 The bias is consistent within the genome

In order to demonstrate the consistency in this bias, the SL117 and SL523 datasets were split into two, assigning each chromosome to one or other subset such that the two subsets contained data from an approximately equal number of nucleotide positions in the genome. The sequence bias for a fragment start occurring in the middle of any 8-mer sequence was calculated for each 8-mer in both subsets where the total number of the 8-mer in the genome exceeded 50,000 (Section 2.2.7). The values were plotted against each other (Figure 2-9). The results demonstrate extremely significant correlation, in that applying the Fisher transformation [33] to the SL523 data to determine the likelihood of such a result if the results from the half genomes were uncorrelated gives a z value of 339, and consequently a p-value that is of the order of 10^{-25000} (Appendix F-2).

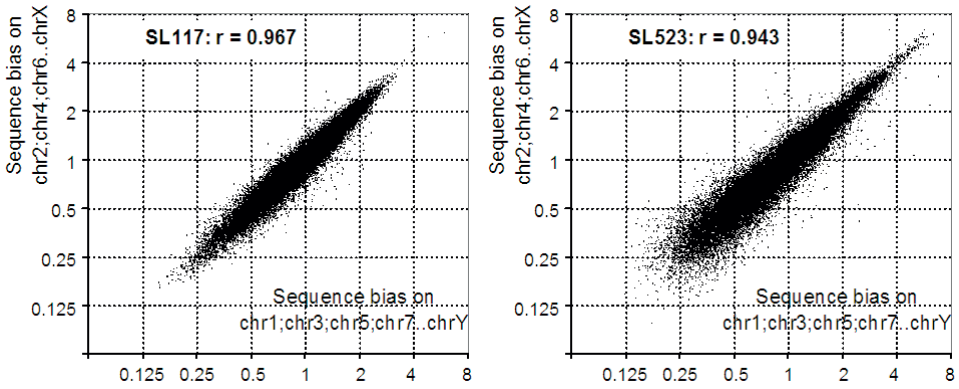


Figure 2-9 Consistency of strong sequence bias within the genome. a) The sequence bias is consistent across the genome as seen in the correlation between the values of Y_s values calculated for each half genome.

This shows that the influence of the DNA sequence on the distribution of fragment start locations is extremely significant. Similar results are seen with the other datasets (data not shown).

2.3.4 Sequence bias varies with nucleotide position and experiment

An approach based on the use of the mutual information between the sequence and the probability of a DNA fragment was used to calculate a set of values that indicate the contribution of each nucleotide to the sequence bias. The interaction intensity, derived from the geometric mean of these values gives an indication of the strength of the sequence bias at a particular position. The spread, calculated using the standard deviation of the log values indicates the degree of interaction between the nucleotide and the other nucleotides when determining sequence bias (Section 2.2.4).

The interaction intensity for the SL117 (Figure 2-10a) and SL523 (Figure 2-10b) datasets are both significantly different from the value of 0.066 obtained with a uniform fragment distribution. There is also a significant variation of the interaction intensity with nucleotide position within the 8-mer. Larger values of the interaction intensity indicate that the nucleotide biases are informative, i.e. a change in the type of nucleotide has a significant effect on the probability of there being a fragment start associated with the sequence. In dataset SL117 the results show that the nucleotides immediately flanking the fragment start location and also unexpectedly at the nucleotide positioned 4th from the fragment start (Figure 2-10a) have a greater role in determining the probability of a fragment start. By contrast in dataset SL523 it is primarily the first two nucleotides of the fragment that have the greatest

significance, with the significance dropping quickly with distance away from the start of the fragment (Figure 2-10b).

The spreads for the SL523 and the SL117 data also both show significant variation across the eight nucleotides, with a peak in the region of the second and third nucleotide from the fragment start.

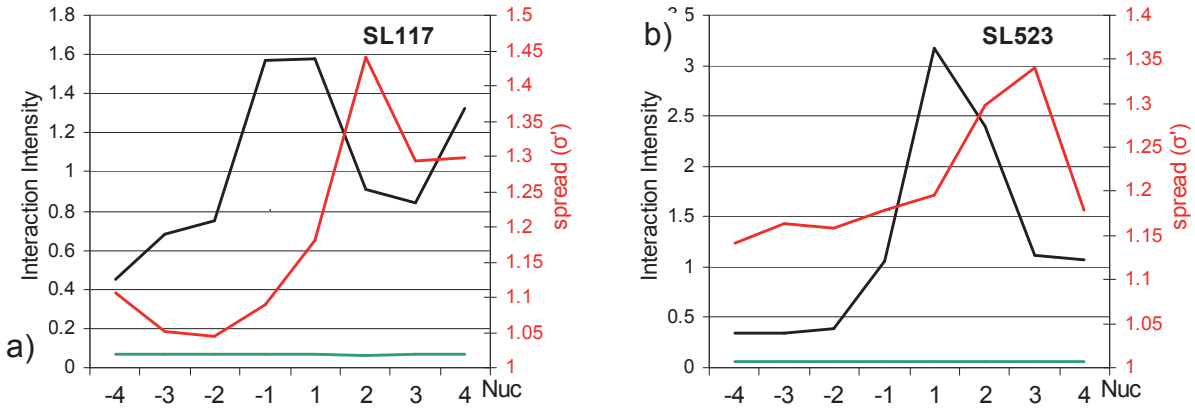


Figure 2-10 The bias of individual nucleotide positions is significantly different from that of uniformly distributed fragments. The interaction intensity gives a measure of how much each of the four nucleotides on either side of fragment start influence the probability of a fragment start being found. X-axis: positive numbers for nucleotides in the fragment, negative for nucleotides immediately preceding the fragment. The interaction intensity is shown as a black line, with the result for a uniform fragment distribution in green. The spread σ', or variation of the interaction intensity with the other nucleotides in the 8-mer is shown in red.

Appendix A shows results from the analysis of 22 other datasets. These show a significant degree of variation of both the pattern of interaction intensity and spread between the datasets.

These results show that there is a significant variation between experiments in the way that individual nucleotides determine the probability of a fragment start occurring. The variation between the technical replicates SL523 and SL117 was also seen when the results from other pairs of technical replicates were investigated.

2.3.5 PCMs and model fitting show significant bias differences between experiments †

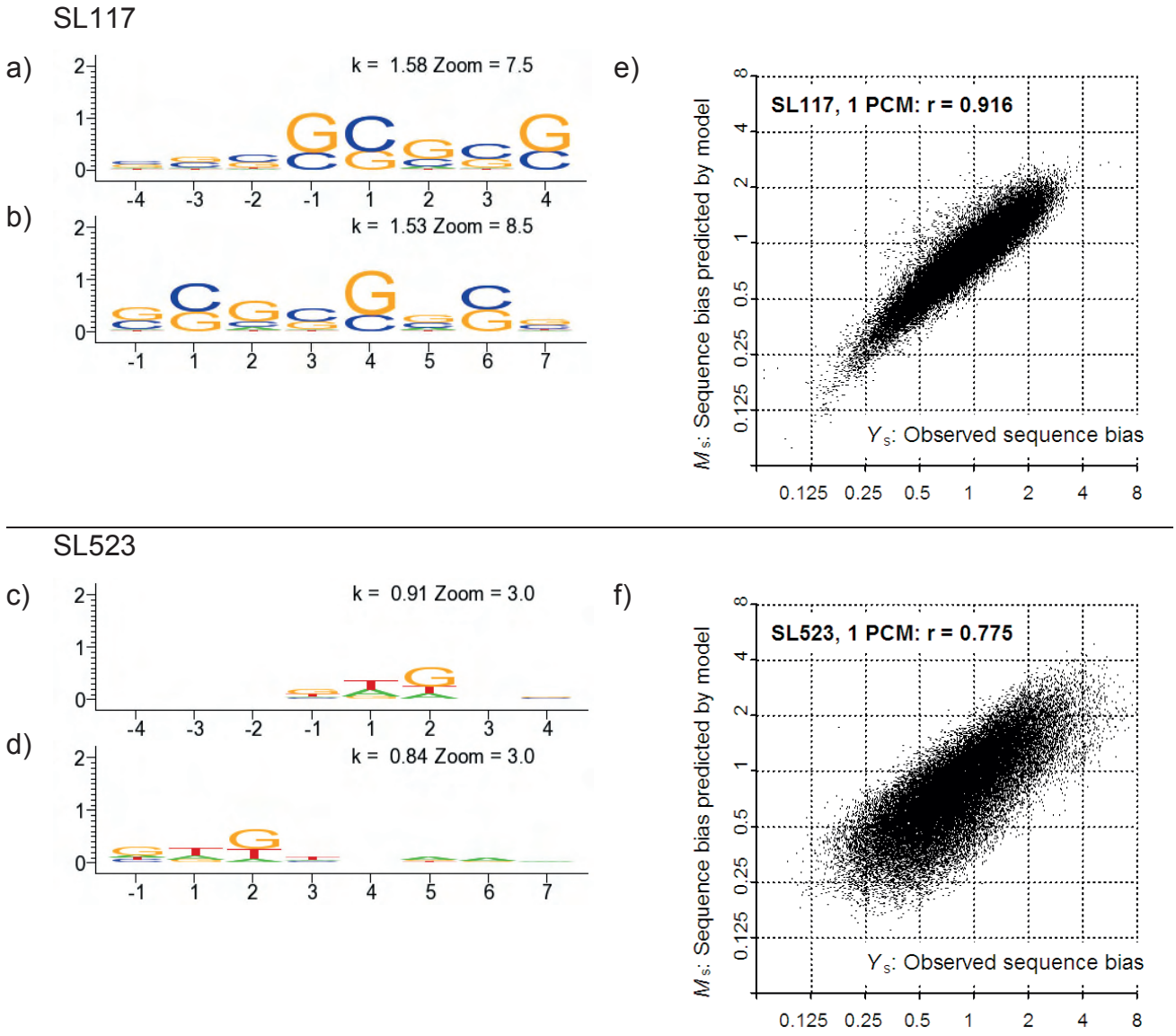


Figure 2-11 SL117 and SL523 PCMs show very different sequence biases. a/c) Sequence biases for four nucleotides on either side of fragment start that gives best fit between model and observed data. k is the global scaling model parameter. The overall height of the column represents the information content of the four coefficients, and the height of each letter indicates the relative values for each nucleotide. "Zoom" is degree of magnification of the information content to make the letters visible in the logo b/d) As a/c but for the 8-mer starting one position before the start of the fragment. e/f) x-y plots show fit between experimental under- and overrepresentation of 8-mers and models shown in b) and d). The offset parameter *o* was only used in the creation of coefficient sets a) and b).

In order to investigate the relationship further, a model was created based on using a Position Coefficient Matrix (PCM), similar to the Position Specific Scoring Matrix (PSSM),

to characterise the bias at each nucleotide position. Model fitting was then used to obtain the closest match between the model prediction and the observed data.

The resulting PCM is represented as a sequence logo, using the same approach as is used to represent PSSM values [87]. In this context the height of each letter is an indication of the degree to which the nucleotide increases the probability of an associated fragment occurring and the overall height of all the letters indicates the degree to which the nucleotide position determines the probability of an associated fragment.

The results of the model-fitting are shown in Figure 2-11a) and c). It can be seen that significance of each nucleotide position as indicated by the overall height of the letters closely matches the profile of the mean information content in 2-10a) and b) respectively.

The model shows that SL117 fragments are more likely to begin between a nucleotide pair consisting of Gs and Cs. The fragments are also more likely to come from locations where at least the first four nucleotides of the fragment, particularly the 4^{th}, will be Gs and Cs. The SL523 data has a very different characteristic, with fragments more likely to come from locations in the genome that result in Gs, Ts and possibly As in the first two positions of the fragment.

The SL117 results suggest that nucleotides further into the fragment may also play a significant role in determining the probability of fragmentation. The analysis method used makes it difficult to consider sequences that are longer than eight nucleotides because increasing the length results in fewer data points associated with each of the possible sequences and therefore greater uncertainty associated with any derived data. In order to assess the significance of nucleotides further along the fragment, the analysis window was shifted to cover only one nucleotide prior to the fragment start and the first seven nucleotides of the fragment. The resulting PCM for SL117 reproduces the CG bias seen at the start of the fragment including the increased bias at nucleotide four and shows that it extends further into the fragment, with an increased GC bias on nucleotide six as well (Figure 2-11b). The PCM for SL523 shows a slight bias towards A at nucleotide positions five and six (Figure 2-11d).

The correlation between the observed and modelled sequence bias indicates that the model is able to reproduce the range of bias seen through the full range of 8-mer sequences (Figure 2-11e and f). The modelling results are consistent with the mutual information based statistics, but provide a more detailed picture of the varying characteristics of different datasets.

2.3.6 Multiple alternative biases exist within each dataset †

The mutual information content analysis shows that the first nucleotide after the fragment break position is the most informative nucleotide in the SL523 dataset (Figure 2-10b). This is not reflected in the model using a single PCM (Figure 2-11c and d) in which the second nucleotide is most significant. This difference could be explained if not a single nucleotide sequence pattern was underlying the sequence bias but a mixture of sequence patterns. The mean information content shows the overall contribution of a nucleotide position in the context of all 8-mers to which each alternative in the mixture would contribute. When represented as a PCM, conflicting alternatives from the mixture could cancel out if there was no net bias to a specific nucleotide.

In order to investigate this possibility, additional PCMs were introduced into the model, representing the possibility that there may be multiple alternative patterns of bias in the nucleotide sequence (Method: Section 2.2.7).

While adding additional PCMs to the SL117 data only made a marginal improvement of the Pearson coefficient from 0.935 to 0.963, adding additional PCMs to the model for SL523 made a significant improvement, increasing the Pearson coefficient from 0.810 to 0.944. In this dataset the range of alternative PCMs have significantly different biases associated with the first nucleotide (Figure 2-12a). This is consistent with the possibility that little bias was seen when a single PCM was used because these alternatives cancelled each other out when represented as the single PCM that can only indicate an overall average.

Section 2.4.3 examines in more detail the problem of choosing how many additional PCMs should be added and concludes that techniques such as using the Bayesian Information Criteria [89] are not well fitted to this problem. The selection was instead based on requiring a degree of variation between all of the PCMs.

An analysis of further datasets from same source, the Myers/HudsonAlpha lab (Appendix C-2 to C-3) shows that the PCMs generated from these data matches the character of either the single PCM SL117 or the multi-PCM SL523 dataset. In the cases where a multi-PCM model is found to be appropriate there is considerable variation in the characteristics of the PCMs between datasets.

Appendix C-4 provides a similar analysis of data from the Yale/UC-Davis/Harvard lab and Appendix C-5 shows four results from *C. elegans*, where it can be seen that the data exhibits slightly different characteristics from the Myers lab data. Some of the results such as Y864-2 (Figure 2-12b) are similar to the SL523 Myers example (Figure 2-12a).

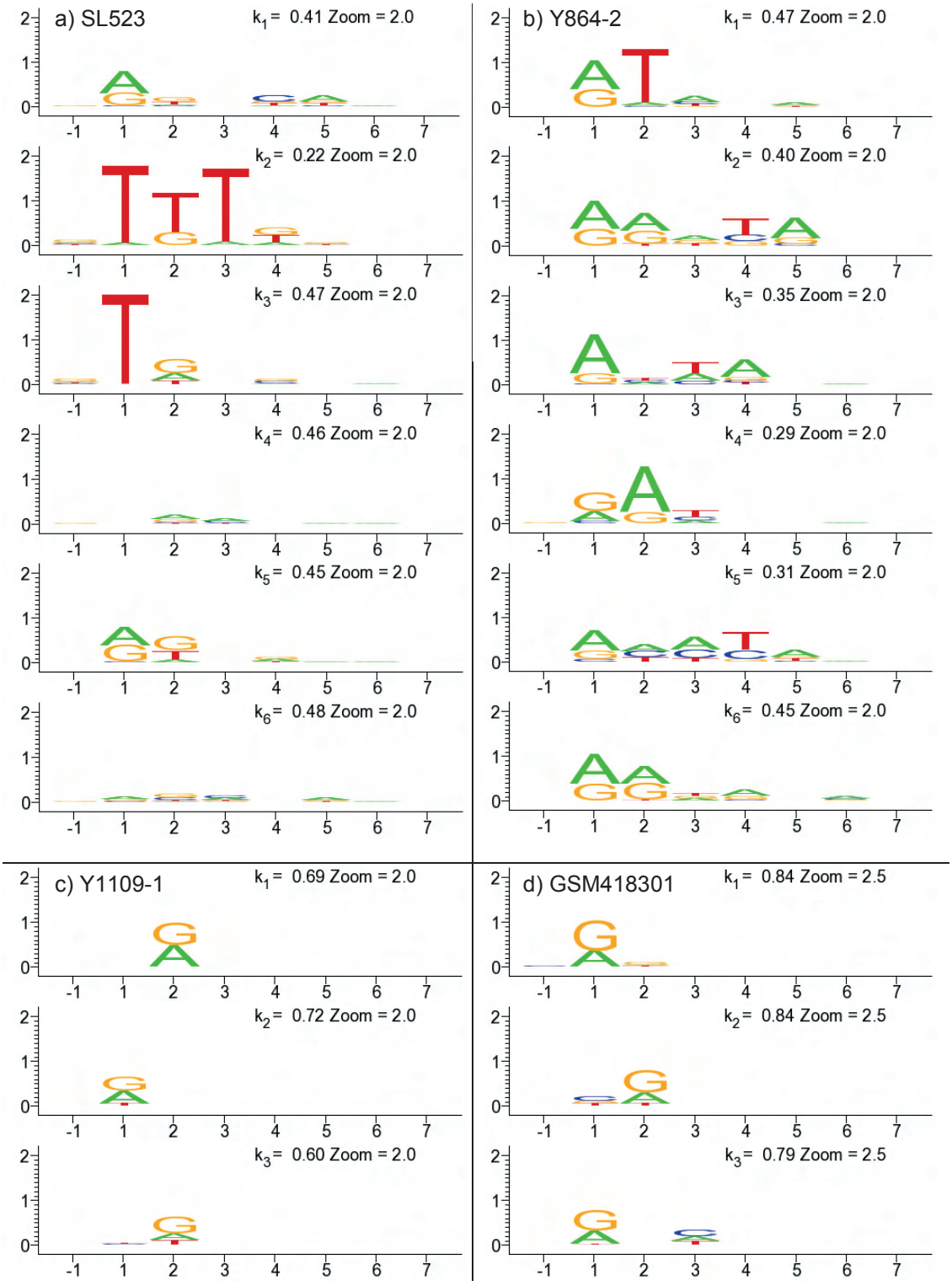


Figure 2-12 ChIP-seq sequence bias PCMs with sequence biases predominantly within the fragment a) shows the difference between a multiple PCM fit and the single PCM fit (Figure 2-11d) of the same SL523 data. b), c) and d) show other datasets where multiple PCMs significant variation in nucleotide distributions were generated by model fitting.

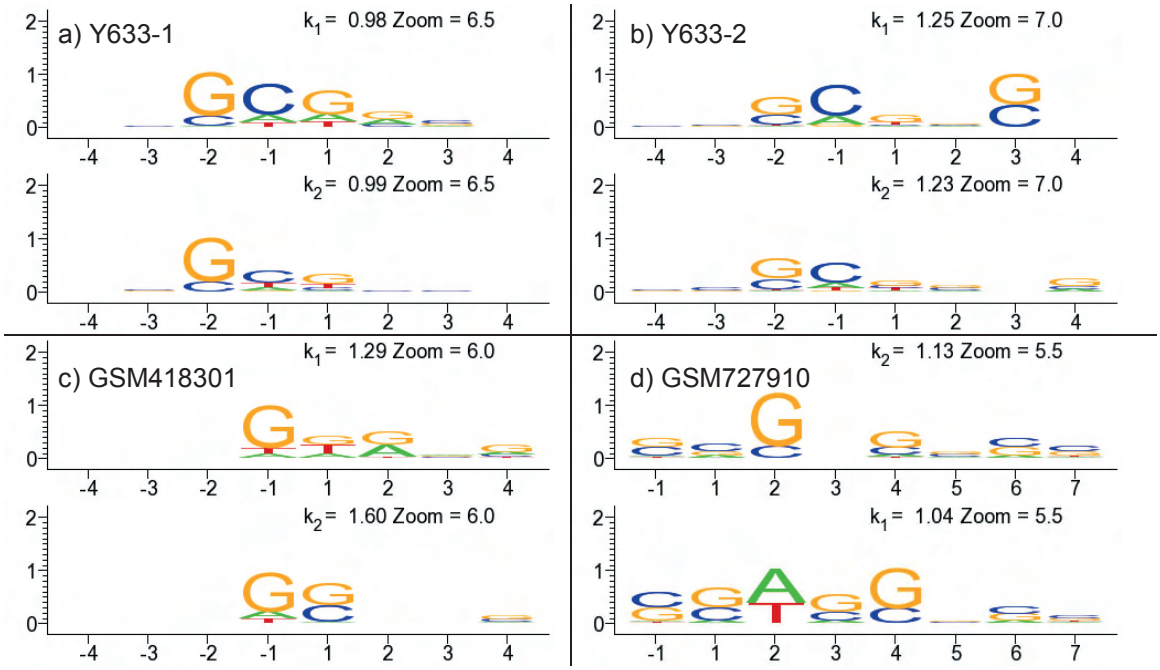


Figure 2-13 ChIP-seq biases with significant bias on both sides of fragment start. As well as significant biases before the fragment start they also show a preference for GC bias. a) shows a dominant bias two nucleotides before the fragment start b) c) and d) all show some degree of additional bias on the fragment side of the break. b) and d) show biases on a specific nucleotide position, which is similar to Figure 2-24. c) and d) show additional fragment side bias that includes non GC nucleotides,

The Y633-1 and Y633-2 PCMs show some similarity to the SL117 PCMs in that they are GC dominant, and with bias both before and after the fragment start position (Figure 2-13a and b). Unlike SL117 they show significant bias up to two nucleotides before the start of the fragment rather than just one nucleotide. There are also examples that appear to show a combination of the two characteristics, with a GC-rich bias component that can be seen on both sides of the fragment position, together with a bias involving all nucleotide types that is associated just with the fragment side of the start position (Figure 2-13c and d).

2.3.7 The information from PCMs is consistent with the identity of over-represented 8-mers ‡

Another view of the relationship between fragments and the genomic sequence can be obtained by looking at the most over represented 8-mer sequences in a particular dataset.

Y1109-1 and Y864-2 in Figure 2-12 provide good examples of PCMs with different characteristics. Figure 2-14 shows the most over-represented sequences in each of these datasets, which give some indication as to how these different PCMs arise. The Y1109-1 data shows that four of the sequences most likely to be associated with fragment starts include a

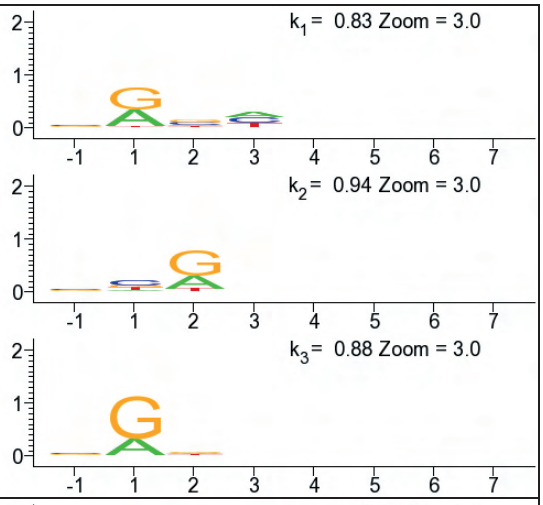
repeated TGGAA motif shifted by varying degrees. The remaining two sequences are the reverse complement of this repeating motif. The symmetrical emphasis on As and Gs in the sequences would appear to correspond to the AG pairs seen in the PCMs.

The Y864-2 data are different, with all the sequences staring with a GATATAA motif, extending to varying degrees into the fragment. This corresponds to the PCMs in that the emphasis on A, and the significant T in the second position can also be seen in the PCMs.

The sequences extending further into the fragment that are associated with the over-represented 8-mer TGGATGG in Y1109-1 were then studied in more detail.

a) Y1109-1

Sequence	Instances	Breaks	Expected	Y_s
TGGAATGG	122974	2306	443	5.19
GGAATGGA	130130	2151	469	4.57
ATGGAATG	141534	2159	510	4.22
GAATGGAA	177851	2382	641	3.70
CCATTCCA	122228	1538	441	3.48
CATTCCAT	141722	1646	511	3.21



b) Y864-2

Sequence	Instances	Breaks	Expected	Y_s
GATATACA	94779	1864	337	5.62
GATATAAT	130174	2538	462	5.57
GATATAAA	178471	3451	634	5.52
GATATAAC	54964	1054	195	5.48
GAATGGAA	177851	3388	632	5.44
GATAACAT	93451	1763	332	5.39

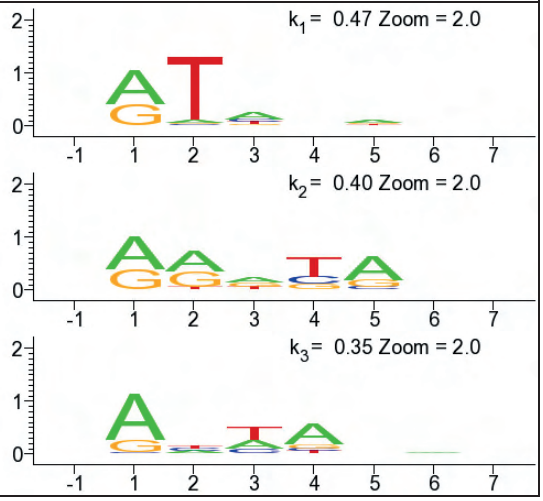


Figure 2-14 There are clear connections between the overrepresented sequences and PCMs The top six overrepresented PCMs in two different experiments (Y1109-1 and Y864-2) are shown together with three of the PCMs resulting from model fitting. Nucleotides associated with common overrepresented motifs are highlighted in yellow. Sequences that occur fewer than 50000 times in the genome have not been included.

a) Y1109-1 Sequence	Fragments	Sequences in chromosomes	Fragments per sequence
TGGAATGGAATCAACTCGAT	22	1	22.00
TGGAATGGAATGGATCAACC	19	1	19.00
TGGAATGGACCAGAATGCAA	14	1	14.00
TGGAATGGAAGGCAATAGAA	10	1	10.00
TGGAATGGACTGGAGAGGAA	21	4	5.25
TGGAATGGAATGGAAAGAAT	29	6	4.83
TGGAATGGAATTTAGTGGAA	9	2	4.50
TGGAATGGGATGGGATGGAA	6	2	3.00
TGGAATGGAAAGGACTGGAA	17	6	2.83
TGGAATGGAATAATCCATGG	12	9	1.33
TGGAATGGACTGGAACAAAA	15	15	1.00
TGGAATGGAATGGAATAATC	7	7	1.00
TGGAATGGAATGTAACGGAA	5	6	0.83
TGGAATGGAGTGAAATGGAA	6	8	0.75
TGGAATGGAATCGAACGGAA	5	7	0.71
TGGAATGGAATGGAGTGGAT	5	11	0.45
TGGAATGGAACAGAATGGAA	8	21	0.38
TGGAATGGCATGGAATGGAA	10	30	0.33
TGGAATGGACTGGAGTGGAA	14	44	0.32
TGGAATGGAATGGAATGGAT	28	110	0.25
TGGAATGGAATGGAATGTAC	5	24	0.21
TGGAATGGAATGGAATGGAG	29	148	0.20
TGGAATGGAATGGAATGGGA	6	34	0.18
TGGAATGGAACGGAATGGAA	13	79	0.16
TGGAATGGAATGGAATCAAC	10	61	0.16
TGGAATGGAGTGGAATGGAA	20	172	0.12
TGGAATGGACTGGAATGGAA	15	130	0.12
TGGAATGGAATGGAAGGGAA	5	43	0.12
TGGAATGGAATGGAACGGAA	12	106	0.11
TGGAATGGTATGGAATGGAA	5	46	0.11
TGGAATGGAATCAAATGGAA	10	117	0.09
TGGAATGGAATGAAATGGAA	10	126	0.08
TGGAATGGAATGGAATGAAA	7	117	0.06
TGGAATGGAATCGAATGGAA	11	228	0.05
TGGAATGGAATTGAATGGAA	6	162	0.04
TGGAATGGAATGGAATGGAA	83	2825	0.03
TGGAATGGAATGGAAAGGAA	5	188	0.03
TGGAATGGACTCGAATGGAA	8	441	0.02
TGGAATGGAATGGAATGGAC	5	343	0.01
Consensus TGGAATGGAATGGAATGGAA			

b) SL117 Sequence	Fragments	Sequences in chromosomes	Fragments/ sequence
TGGAATGGACCAGAATGCAA	7	1	7.00
TGGAATGGAATCAACTCGAT	5	1	5.00
TGGAATGGACTGGAGAGGAA	6	4	1.50
TGGAATGGAATGGAAAGAAT	5	6	0.83
TGGAATGGAAAGGACTGGAA	5	6	0.83
TGGAATGGACTGGAACAAAA	5	15	0.33
TGGAATGGCATCGAATGGAA	5	35	0.14
TGGAATGGAATGGAATCAAC	5	61	0.08
TGGAATGGAATGGAATGGAT	5	110	0.05
TGGAATGGACTGGAATGGAA	5	130	0.04
TGGAATGGAATGGAATGGAA	5	2825	0.00

Table 2-2 All fragments associated with the TGGAATGG 8-mer from two datasets.

a) Y1109-1 where there was a significant break bias associated with this 8-mer b) SL117 where it was no significant bias. In each case the number of instances of the 20-mer in the genome is shown, together with the number of fragments seen that start with the 20-mer. The over-represented fragments always start with variants of a repetitive sequence involving the 5-mer TGGAA

Table 2-2a) lists all the fragments reads from the Y1109-1 dataset associated with this sequence where there were two or more instances in the forward strand across the whole genome. Table 2-2b) shows the equivalent data from SL117 where the normalised sequence

bias for TGGATGG was only 1.8, compared to 5.19 in Y1109-1. This shows that there are significantly fewer instances of the sequence, which is consistent with the lower bias for this 8-mer in the SL117 dataset.

The consensus across the 20 nucleotides in both sets of data is that the TGGAA motif repeats up to four times, with decreasing accuracy at greater distances from the fragment start.

2.3.8 Sequence bias from immunoprecipitated fragments and input DNA is poorly correlated

The two datasets SL116 and SL522 are partners to the SL117 and SL523 datasets in that SL116 and SL522 are from fragments which were immunoprecipitated to select for the NRSF protein whereas SL117 and SL523 are the input data produced at the same time from the same cell lines and intended as controls. Immunoprecipitation results in fragments that are clustered around the locations where the target proteins bind to the DNA, and it is this clustering that is used to determine the location of the NRSF binding sites (Section 1.4.9).

The PCMs obtained from model fitting input DNA and immunoprecipitated data from the experiments on the same cell line have similar, but not identical, characteristics (Figure 2-15). This raises the question as to the significance of these differences (particularly in the later SL522/523 datasets) and the origin of any differences.

One way of studying the differences is by looking at the correlation between the two datasets of the sequence bias for the full set of 8-mers. Such a comparison shows that there is very poor correlation between the SL522 and SL523 datasets despite the apparent similarity in the PCMs created (Figure 2-16a).

Equation (2.5) assumes that the background fragment distribution is uniformly distributed around the genome, and no attempt has been made to adjust the equation to allow for the clustering of immunoprecipitated fragments, so one possible explanation for the poor correlation may be as a result of the clustering in the immunoprecipitated SL522 data. One way of trying to adjust for this effect is to look at the statistics of fragments excluding those that come from the peaks. The results from the SL522 data shows that there is good correlation between the original data and the data with the peaks removed (Figure 2-16b) suggesting that the effect of the greater fragment density in the peaks does not appear to be sufficient to explain the lack of correlation in Figure 2-16a.

The effect of removing the peaks (Figure 2-16b) will be that the sequence bias of those sequences that are found in the peaks will be reduced, giving rise to the points that fall below the 45 degree line. There will be little change in the bias for those sequences that are found

seldom if at all within the peaks, giving rise to the clear line at an angle of 45º in this figure. The results indicate that the sequences found within the peaks are a mixture of sequences with under and over-represented fragments.

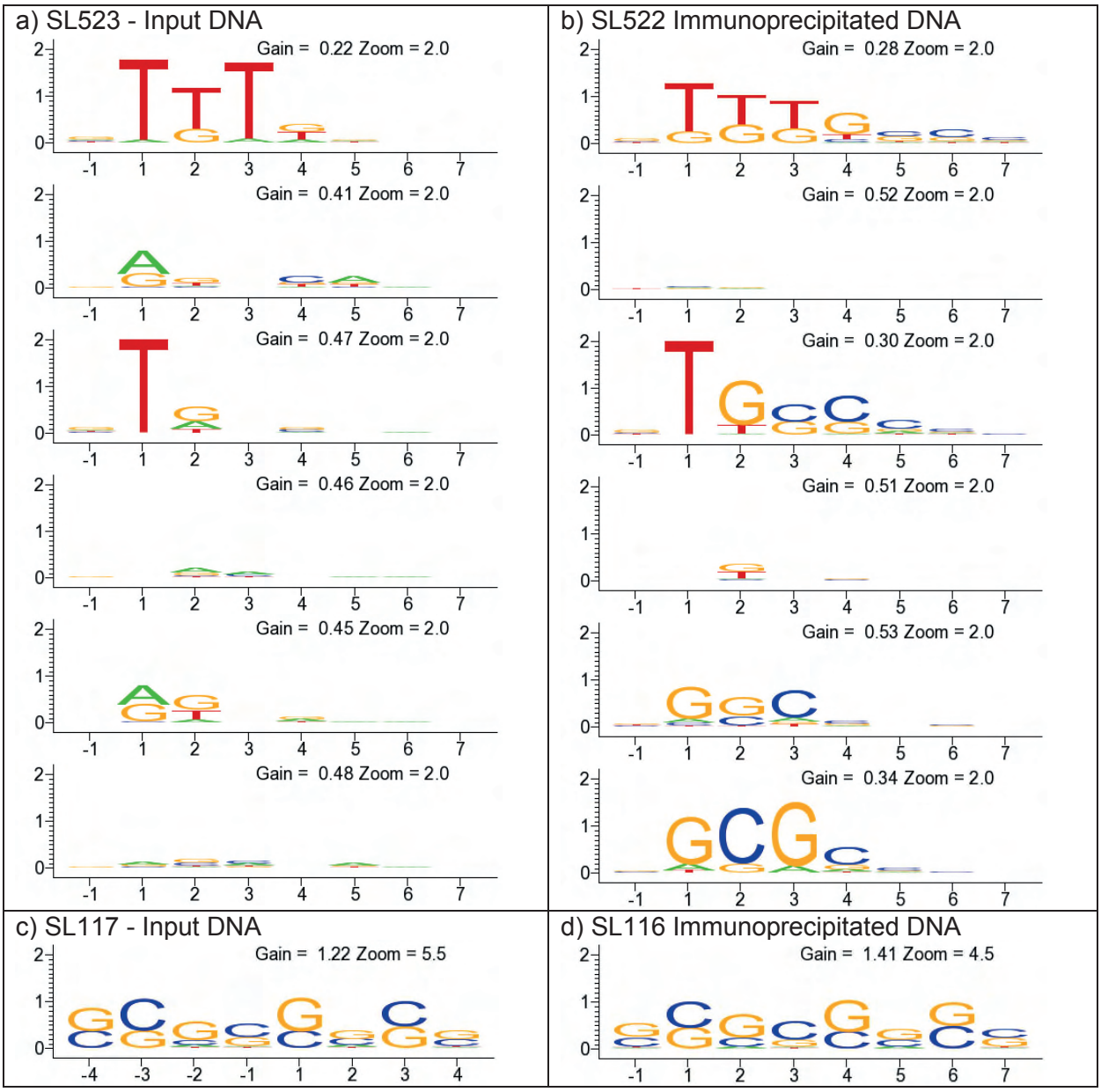


Figure 2-15 Similar PCMs from model fitting input and immunoprecipitated data. Model fitting of the sequence bias of immunoprecipitated fragments in b) and d) show similar characteristic PCMs to those obtained from model fitting input data from experiments conducted at the same time with the same cell line a) and c).

There is however very poor correlation between the peak-excluded SL522 data SL523 data (Figure 2-16c). This suggests that the lack of correlation between the two datasets is probably not as a result of the clustering in the SL523 data but is instead as a result of a more

fundamental difference between the two sets of data such as the likelihood that the data were from two different experiments.

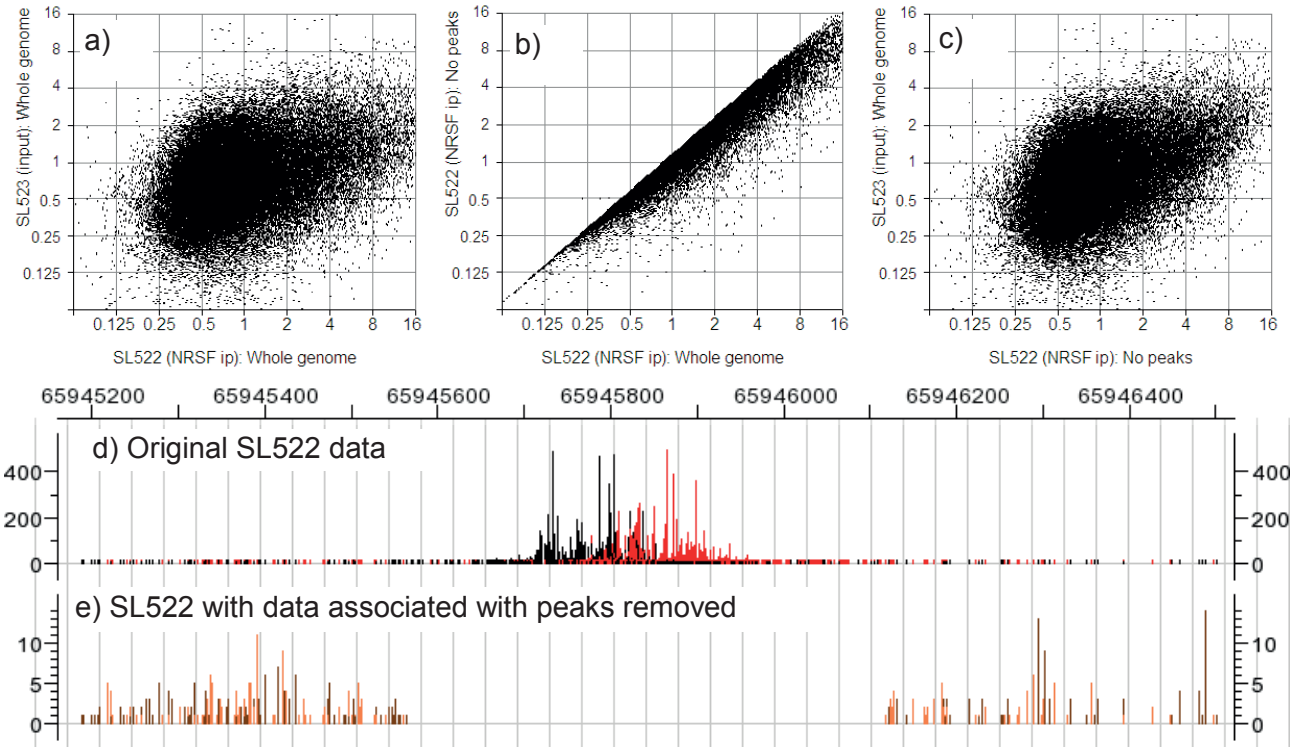


Figure 2-16 Sequence bias of immunoprecipitated DNA is poorly correlated with the input DNA sequence bias. a) There is very poor correlation between the sequence bias of immunoprecipitated and input fragments from the same cell line that were published as a control. b) Correlation between SL522 sequence bias with and without data from peaks showing good correlation. c) Poor correlation between peak-excluded immunoprecipitated data and input data c). d/e) A sample of SL522 with and without peak fragments. b) and c) suggest that the poor correlation in a) is not because of bias introduced by having significant numbers of fragments from a small number of regions in the genome, but instead from an underlying lack of correlation between the two datasets.

2.3.9 Datasets with different PCMs also show different fragment distributions

The earlier analysis in this chapter was largely concerned with the input fragments. However, the fragment distribution of input DNA is sufficiently sparse that it is not easy to distinguish any clear pattern in the fragment distribution along the genome that might be related to the sequence bias. For example, the distributions in Figure 2-7 come from datasets that have very different sequence biases, and yet while there does appear to be some differences in the distribution; the sparseness makes it difficult to discern any specific way in which they differ.

Immunoprecipitated fragments however, have tight clusters of fragments making it possible to see how sequence bias might affect the fragment distribution within the clusters.

The SL116 and SL522 data considered in the previous section are from immunoprecipitated fragments which are representative of the two different types of sequence bias that is seen in the Myers/HudsonAlpha lab data (Figure 2-15b and d). The peaks in the SL116 and SL522 data, although containing a similar number of fragments, show very different fragment distributions (Figure 2-17a and b). There is a much greater variation in tag counts between adjacent nucleotides in the SL522 data than there is in the SL116 data. However, if a rolling average with a 50 bp window is used, which is similar to the approach used in most peak finding algorithms, the overall fragment distribution of the fragments within the peaks is very similar (Figure 2-17c and d).

The similarities of the averaged data will mean that peak finding algorithms will tend to come to similar conclusions about the location of the peak as they use a similar smoothed version of the data (Section 1.4.9).

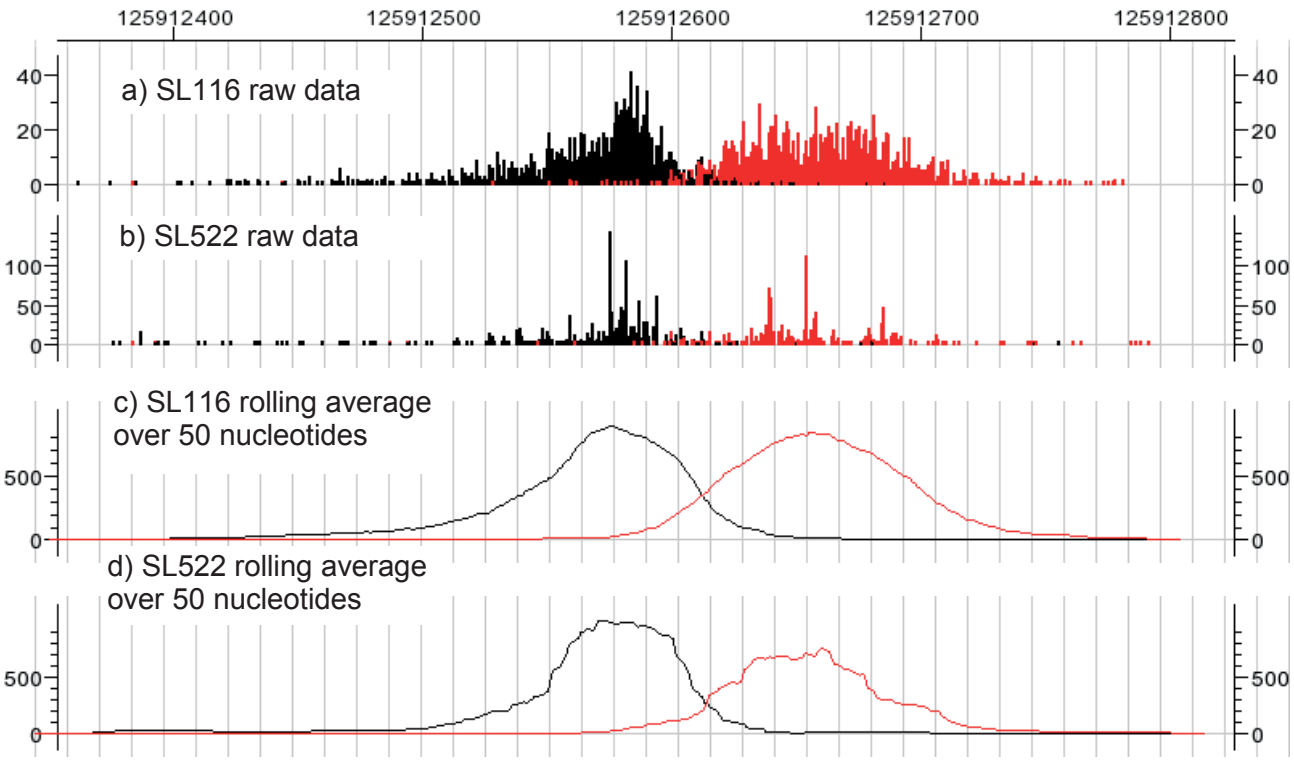


Figure 2-17 Similarity of rolling average of tag distribution at binding peaks contrasts with significant differences in underlying tag distribution Graphs show nucleotides 125912360-125912819 of Chromosome 9. a/b) Histogram showing the number of fragments starting at every location c/d) As a/b) but using a rolling average with a window size of 50 nucleotides.

2.3.10 Adjustment of fragment distribution for sequence bias

The quantification of the sequence bias within a dataset makes it possible to adjust for this bias. This section investigates some of the ways in which such an adjustment could be made.

Results

While the SL522/116 data has different fragment distributions in the regions of the peaks, it is nevertheless unclear to what extent this is related to the different sequence bias in the two sets of data. In order to investigate this further, the fragment distribution for SL522, which shows the greatest fragment density variation, was adjusted to compensate for the effects of the sequence bias (Methods section 2.2.10).

It had been thought that it would be possible do this compensation using sequence bias information from the input datasets, on the assumption that this would have the same underlying bias as the immunoprecipitated data. However Section 2.3.8 shows that the bias characteristics for SL523, the apparent input dataset for SL522 is very different from SL522, making it unsuitable as the source data for performing such a normalisation.

As a result, the SL522 data itself was the only option for a reference source that could be used for performing the bias correction. While the sequence bias for the whole genome data is well correlated with the data that excludes the peaks (Figure 2-16b), the sequence bias data derived from SL522 data from which the data from the peaks had been excluded was used. This reduces the degree to which sequence bias that is specific to the peaks is used in the normalisation as the intention is only to compensate general bias characteristics seen in the data. This means that the data in Figure 2-16e rather than 2-16d was used as the reference data for bias compensation.

Once a dataset has been chosen or derived to be used as the reference for the normalisation, the simplest approach is to use the bias for all of the 8-mers in the reference data to perform the bias correction (Figure 2-18b).

A problem with this approach is that the number of data points for some 8-mers may be very small, such that there will be significant variance in the measured sequence bias as a result of the small sample size. To demonstrate this with some examples; in the SL522 reference dataset there are 38 8-mers where there are no instances of associated breaks. The normalisation multiplier for such 8-mers would be ∞, which would cause a problem if there were any of these 8-mers in the data to be normalised.

There are a further 109 8-mers in the reference data for which there is only one associated fragment. If the random variation in the process had resulted in one more or fewer fragments being associated with any of these 8-mers then the change in the normalisation factor associated with them would be dramatic.

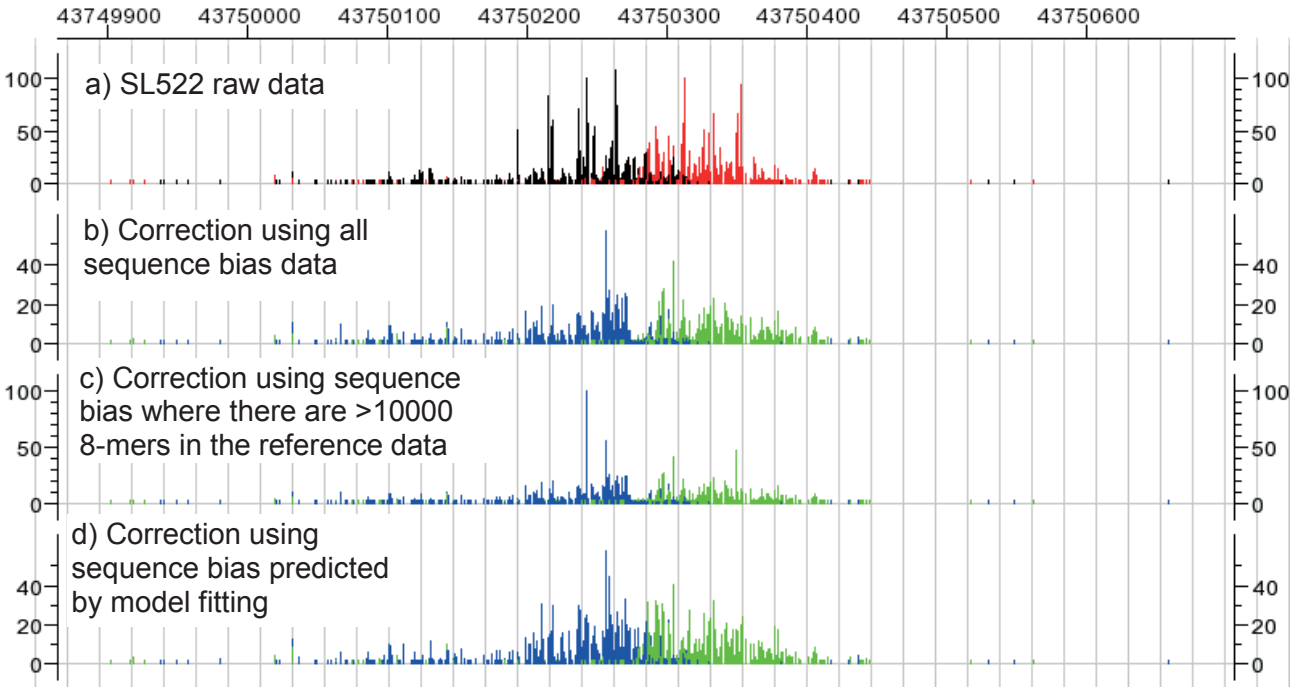


Figure 2-18 Correction for sequence bias reduces some of the noise in ChIP-seq peaks. Chromosome 22: 43749875 to 43750694: SL522. Forward reads shown in Black & Blue, Reverse reads Red & Green. a) Raw ChIP-seq peak. b) Distribution after correction using all of the sequence bias information from fragments not associated with peaks c) Correction only using sequence bias where there are at least 10000 instances of the sequence in the genome leaves some significant spikes. d) Correction using the sequence bias predicted by model fitting.

One approach to overcoming this problem is to apply a weighting that is a function of the number of 8-mers in the reference data, which would scale back the degree of normalisation when there are fewer instances of the sequences. A simplistic version of such an approach is simply not to rescale locations where the associated 8-mer occurs fewer times in the genome than some arbitrary threshold, such as 10000. This threshold was selected because all of the 8-mers for which there are only zero or one associated breaks that occur less than 8000 times (Figure 2-18c). This threshold value excludes approximately 20% of the 8-mers. However these are, by definition, the 8-mers that occur least frequently in the genome and this threshold only affects the bias compensation for 1.7% of the locations in the genome.

A third approach is to use the sequence bias that is predicted by the model after the parameters have been set by model fitting to the reference data. This approach avoids some of

the problems associated with 8-mers with small sample size in that the bias for such 8-mers is derived from the larger collection of the 8-mers that the model fitting has deemed to have similar characteristics (Figure 2-18d). The model fitting can be considered to have performed a degree of averaging of the sampling noise associated with any specific 8-mer for which there are few associated fragments, and can then be used to predict the underlying bias associated with these 8-mers.

Results: Assessment of improvement

In order to assess the improvement if any made by using these normalisation techniques it is necessary to make an assumption about what the data would have been if the underlying DNA sequence had no influence on the probability of fragmentation. This is because any assessment is going to attempt to quantify how much closer looks to this ideal. The complexity of the system makes this difficult when considering any specific region, such as that shown in Figure 2-18, so a simple approximation based on averaging is probably the best that can be acheived.

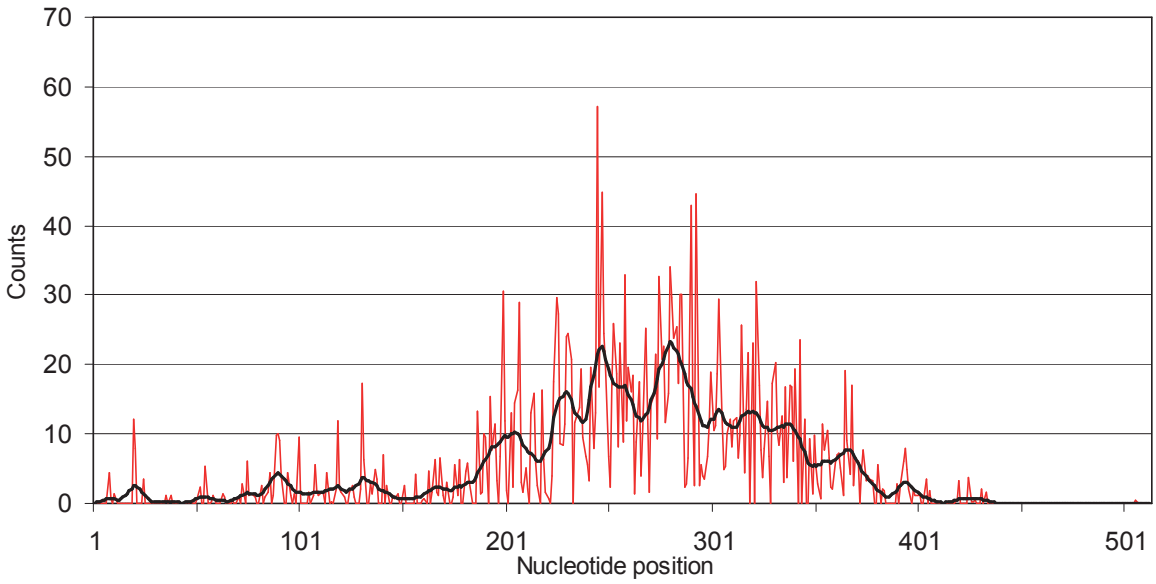


Figure 2-19 Averaging can be used to give an indication of the underlying fragmentation pattern if the DNA sequence does not influence fragmentation. The region is the same as Figure 2-18. The red line is the combined SL522 forward and reverse reads and the black is these data after filtering with a triangularly weighted filter of length 15 nucleotides.

The approach adopted was to create a reference by smoothing the raw data with a triangularly weighted filter of length 15 nucleotides (See also equation(3.2)). The noise that is superimposed on the underlying signal can be considered to be the difference between the raw

and data and the smoothed data. The variance of the raw data would then approximate to the variance of the underlying signal plus the variance of the noise.

Some noise would be expected because each sample in the raw data would be expected to have a Poisson distribution based on the expected value from the underlying distribution. The variances of the raw data, as well as the signal and noise when analysed in this way give an indication of the degree to which the noise is reduced as a result of this normalisation (Table 2-3) for the four sets of data shown in Figure 2-18.

	$\mathrm{Var}(T)$	$\mathrm{Var}(S)$	$\mathrm{Var}(N)$	$\mathrm{Var}(P)$	$\frac{\mathrm{Var}(N)}{\mathrm{Var}(S)}$	$\frac{\mathrm{Var}(P)}{\mathrm{Var}(S)}$
Raw	234.6	91.28	125.24	3.08	137.20%	3.37%
Compensated: all data	40.29	19.09	18.8	1.31	98.48%	6.86%
Compensated: threshold data	62.38	23.29	35.2	1.45	151.14%	6.23%
Compensated: model	67.96	36.05	28.61	1.88	79.36%	5.21%

Table 2-3 Compensation using sequence bias predicted from the model yields most improvement in signal to noise ratio. The variance of the raw data (T) together with the variance of the smoothed data or signal (*S*) and the difference or noise (N) allows an indication of the noise to signal ratio to be calculated. The variance expected assuming the count at each nucleotide position has a Poisson distribution (P) shows that the noise variance greatly exceeds that expected on this basis.

Discussion and conclusion

In all three methods examined for normalising the data to compensate for sequence bias it can be seen that the spikes associated with high number of breaks at the same location are considerably reduced in height. This suggests that the spikes are as a result of non-uniform fragment distribution caused by the sequence bias.

In the data where no normalisation is performed for 8-mers which occur less than 10000 times there is one significant spike that remains (Figure 2-18c). When all of the 8-mers are normalised this spike disappears, indicating that it is associated with an 8-mer that occurs fewer than 10000 times in the genome (Figure 2-18c). There is no evidence that spurious additional spikes are generated when infrequent 8-mers are used to generate bias corrections in either in the data shown in Figure 2-18c or in other peaks that have been examined. This is not unexpected in that such spurious peaks would be associated with 8-mers which occur infrequently and where the break count is low, so it is unlikely that there would be counts associated with such 8-mers within the peaks that would be inappropriately amplified during bias correction.

The estimations of the signal to noise ratio suggest that the residual noise when the thresholded data is used for normalisation is as great as is the case when the raw data is analysed, which will be because of the dominant effect of the small number of large samples that are excluded from the normalisation. The effect of normalisation using either the sequence bias derived from the raw data or the model reduces the noise level relative to the signal by 30% and 40% respectively.

The results suggest that using bias correction derived from all of the 8-mers, or bias correction derived from the modelling are both good candidates as a general technique that could be used. Both of these techniques were used in the investigation described in Chapter 4 where, for the data being investigated, using bias derived from all of the 8-mers seemed to give the best results.

2.3.11 There is a correlation between sequence bias and 8-mer frequency for some datasets

One of the other differences between the characteristics of the SL117 and SL223 datasets is the relationship between the number of each of the 8-mers in the genome and the sequence bias for the 8-mer. In the SL117 dataset, 8-mers that only occur 1,000 times in the genome are 16 times more likely to cleave during sonication than 8-mers that only occur 300,000 times (Figure 2-20a and b). This relationship is not seen in the SL523 data. The SL117 data also appears to show a clustering of the data points into three separate regimes, grouped by the frequency of the 8-mers in the genome and delineated by a region where there is a smaller spread of sequence biases. The same grouping also exists in the SL523 data.

The relationship between bias and the number of 8-mers was also calculated for an artificial dataset where the breaks were uniformly distributed in the genome (Figure 2-20c). The distribution of apparent sequence bias in a uniform distribution would be expected to follow a Poisson distribution (Section 2.3.2) and normal approximation to the Poisson distribution was used to calculate the three standard deviation distance away from expected sequence bias of one, within which the values from a uniform distribution should fall. The results for a uniform distribution fall within the $\pm 3\sigma$ range and show clearly how the large sample size for the sequences that occur frequently in the genome should result in a very small spread in sequence bias if the sequence had no influence on the probability of fragmentation. The SL117/523 results fall well outside this range, showing again the significant effect of the DNA sequence on the likelihood of fragmentation.

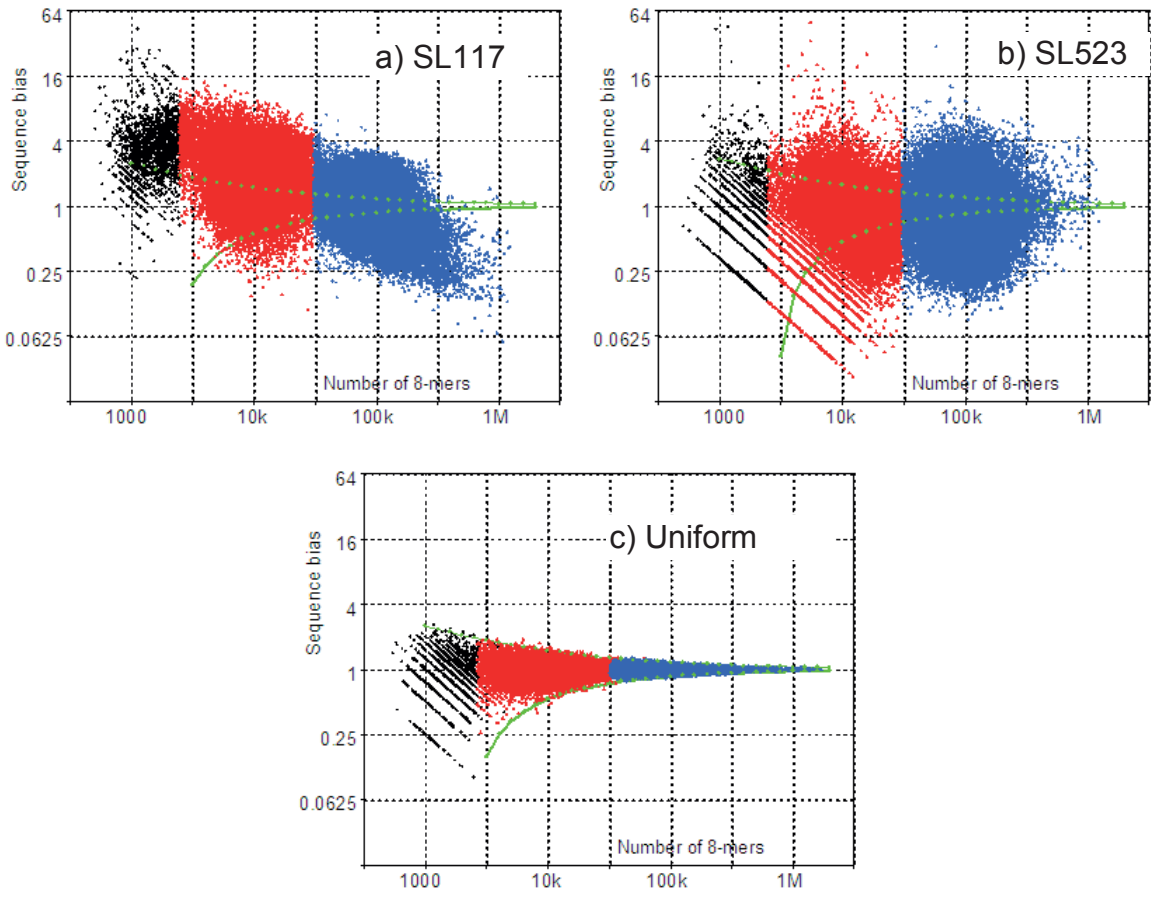


Figure 2-20 In some experiments there is a correlation between the number of 8-mers in the genome and sequence bias. All graphs plot sequence bias against the number of sequences in the genome for all 65536 8-mer sequences. a) SL117 data shows a strong inverse correlation between the number of 8-mers in the genome and the probability of fragmentation. Sequences that occur more frequently are less likely to fragment. b) SL523 data does not show the same relationship. In both cases three apparent data regimes are marked by different coloured datapoints. Straight lines in b) are a consequence of 8-mers with low sequence bias and only a few instances in the genome only having a small number of associated fragments. c) Correlation for the artificial dataset with breaks distributed uniformly throughout the genome. In all cases the green lines indicate the $\pm 3\sigma$ limits derived from a theoretical model within which over 99% of the points should lie if there is no relationship between the sequence and the sequence bias.

An examination of two sets of data from *C. elegans* which have different sequence bias characteristics (Figure C-8) shows a different relationship between 8-mer population in the genome and the sequence bias for the 8-mer compared to that seen in the *H. sapiens* data (Figure 2-21). The results lie well outside the expected range for uniformly distributed breaks, demonstrating again the effect of sequence on the probability of fragmentation. While they do not show the grouping into three regions that was seen in the *H. sapiens* data, one of the

samples also shows a slight tendency for the frequently occurring sequences to have a slightly lower sequence bias.

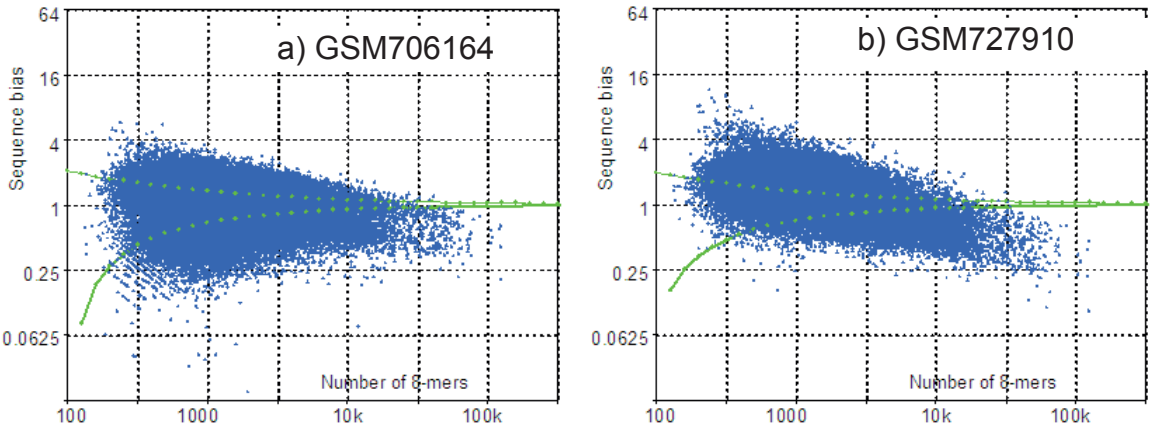


Figure 2-21 *C. elegans* shows a different relationship between sequence bias 8-mer population and sequence bias than that which is shown by *H. sapiens* data. These results show very little variation in the mean sequence bias with the number of 8-mers in the genome. The greater spread in sequence bias for 8-mers that occur infrequently in the genome is as would be expected because of the lower sample sizes.

2.4 Supplementary results

The following results, although significant, are not central to the main thesis but relate more to some of the work that was done to determine and optimise the details of the analysis performed in this chapter.

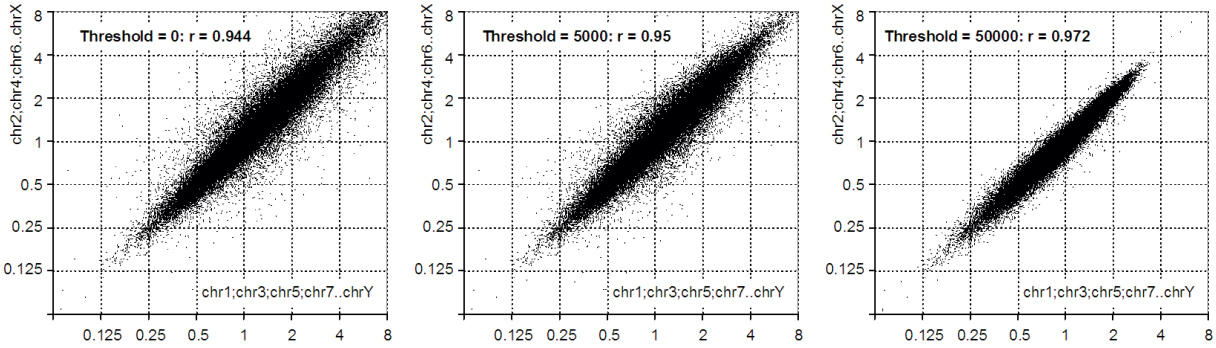
2.4.1 Selecting only a subset of the sequences reduces 'noise' from low sequence counts without introducing systematic errors

Introduction

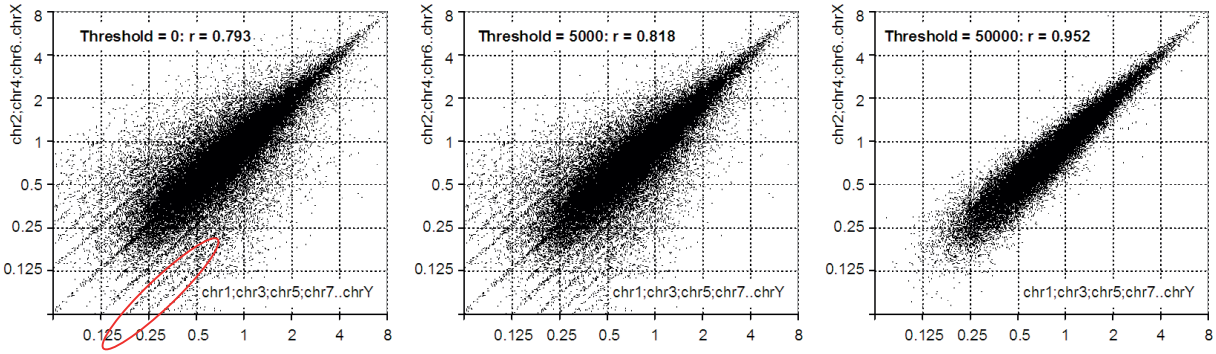
There is a significant variation in the numbers of each of the 65536 different 8-mer sequences in a typical genome. There are potential difficulties with using the data from 8-mers for model-fitting when there are only a few instances in the genome because there may only be one or two fragments associated with these 8-mers in a dataset. This mirrors the problem with using such data for bias correction (Section 2.3.10). The solution adopted to reduce the noise contribution from such points and reduce the computational load during model fitting was only to use sequence data where the number of associated 8-mers exceeded some threshold. This section examines the impact of this approach.

Analysis

a) SL117



b) SL523



c)

Sequence count threshold N	Number of 8mers with more than N instances	Pearson coefficient SL117	Pearson coefficient SL523
0	65536	0.9440	0.7935
5000	61924	0.9504	0.8181
50000	37134	0.9724	0.9522

Figure 2-22 Variation of bias correlation with threshold a) and b) x-y plots showing the correlation of 8-mer bias using data from the two halves of the genome for SL117 and SL523. Each point compares the bias of the same sequence from the two half genomes. Data for sequences that occur fewer times than the three values of threshold shown are excluded from the graphs and calculations. Lines associated with quantisation of break counts are very visible in the SL523 data at low thresholds. Red oval indicates 8-mers where the ratio of the number of breaks in the two half genomes is 2:1. This artefact is not present when a threshold of 50000 is used.

The first 25 nucleotides of each fragment were sequenced in experiment SL117 in order to align the fragments to the genome. A 25 nucleotide sequence is sufficient to identify approximately 4.48 billion unique sequence tag positions in the human genome. In the SL117 dataset there were 19.3 million tags that were able to be uniquely mapped to the genome. If the DNA sequence was essentially random then any given 8-mer would occur approximately 68400 times in the genome, and in the 19.3 million reads there would be expected to be an

average of 19,300,000/65536 = 294 fragments associated with each sequence. The non-random nature of the DNA sequence means that some sequences are significantly underrepresented (For example, CGCGTACG only occurs 503 times in the mappable regions of the human genome) and the number of breaks associated with the sequence is consequently very low (There were only 11 instances where the data shows a break occurs between the C and G at the start of the CGCGTACG sequence). Any data derived from sequences which occur so infrequently will be very noisy.

Figure 2-22c) shows how many of the 65536 different possible 8-mers have more than N instances in the genome for various values of N.

In order to assess the impact of using a subset of the N-mers, the genome was split into two, assigning each chromosome to one or other subset such that the two subsets are approximately equal in size. The sequence bias for each of the 8-mers was calculated for both of the two half genomes and plotted against each other. This was done for the two datasets and for various threshold values (Figure 2-22a and b).

Larger values of N remove the sequences with fewer instances across the genome, and in both datasets this removes the outliers around a central core distribution, reducing the noise associated with the distribution and improving the Pearson correlation coefficient. Horizontal lines in the SL523 data result from 8-mers for which there is a combination of only a few instances of the 8-mer in the genome and also a low sequence bias, resulting in just one or two fragments being associated with the 8-mer.

The quantisation of the results to integral numbers of fragments results in the diagonal lines which are associated with bias ratios that are a ratio of two low integer values. This gives an indication of the types of artefacts that can occur with 8-mers associated with such low fragment counts, raising concerns that this could cause other subtler effects during modelling. Using a threshold of 50,000 removes the points where this artefact was most obvious, without appearing to distort the general distribution of data.

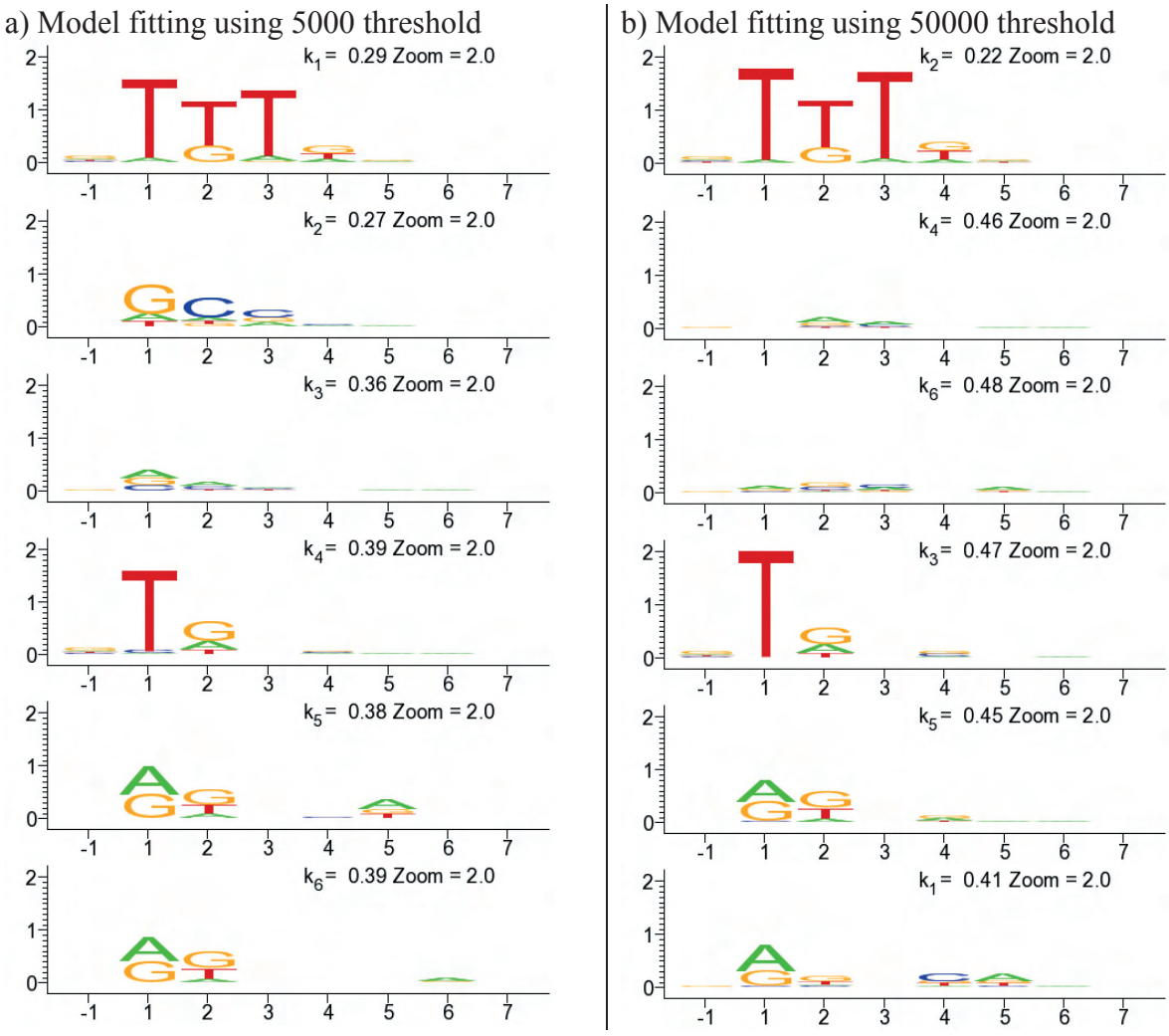


Figure 2-23 Comparison of SL523 PCMs generated by thresholds set to 5000 and 50000. This shows that broadly similar characteristics are obtained with the two different thresholds, although there are some subtle differences. Model fitting with a 5000 threshold results in a PCM with a C at position two, implying that there are a number of over-represented 8-mers with a C at this position but that they tend to be associated with 8-mers with fewer than 50000 instances in the genome.

In order to test for possible effects due to using different thresholds, the model fitting results obtained using two different thresholds were compared. Model fitting was used to generate PCMs for the two datasets with the threshold set to 5000 (which includes 94.8% of the 8-mers) and 50000 (which includes 56.7% of the 8-mers). The two PCMs for the SL117 dataset were essentially identical (Figure 2-23a). The two sets of PCMs for the SL523 dataset were very similar, but showed a very slight difference (Figure 2-23b). Pearson coefficients were used to test the degree of model fit for the SL523 data which also showed that the fit was largely independent of the choice of threshold between 5000 and 50000. (Table 2-4). The results show that the Pearson coefficient is determined predominantly by the threshold used in the evaluation of the PCMs rather than the threshold used to generate the PCMs. Both sets of

results indicate that no systematic errors are introduced as a result of using a threshold of 50000 instances of an 8-mer in the genome when working with data from *H. sapiens*.

	Coefficients optimised with 5000	Coefficients optimised with 50000
Tested with 5000	0.8725	0.8521
Tested with 50000	0.9263	0.9383

Table 2-4 Pearson coefficients indicate equivalence of PCMs generated with different threshold values. PCMs were generated with thresholds set to 5000 and 50000 and then Pearson correlation calculated for the fit between model and observed data for both sets with both thresholds. Values are largely determined by the test conditions and not the threshold used to generate the coefficients.

2.4.2 An offset parameter improves model-fit in single PCM cases

Initial model fitting showed that there was frequently a systematic problem with model fitting in the datasets when the optimal number of PCMs required for model fitting appeared to be one. An example is shown in Figure 2-24a where the best fit line of the correlation between the observed and model fitted data has a slight curve, indicating that the model is unable to fit the sequence bias accurately when the sequence bias is very small.

An additional offset parameter O was added to the model as follows:

$$S_s = O + \max\left(k_j \prod_{i=x}^{y} p_{i,n_i,j} / 4\right)_{j=1}^{P} \tag{2.36}$$

This allows the fitting algorithm to add a constant negative value to all of its bias predictions, which will have a greater significance for the sequences where the model predicts a smaller bias value. This improved the result of model fitting, as shown in Figure 2-24b).

While this does remove the systematic error, because of the small number of sequence values involved, the effect on the Pearson coefficient is marginal (an improvement from 0.923 to 0.928 in the example of 2-24). There was also only a very marginal change to the PCM that was generated.

The poor fit at low sequence biases appears to be because the algorithm for converting from the PCM to sequence bias is non-optimal for very low biases. A brief investigation of some alternatives did not produce an algorithm that gave improved predictions (data not shown).

The problem only arose when the model only incorporated a single PCM, so the offset was only included in the single PCM model when there was evidence of the distinctive curve in the correlation plots between model and observed data.

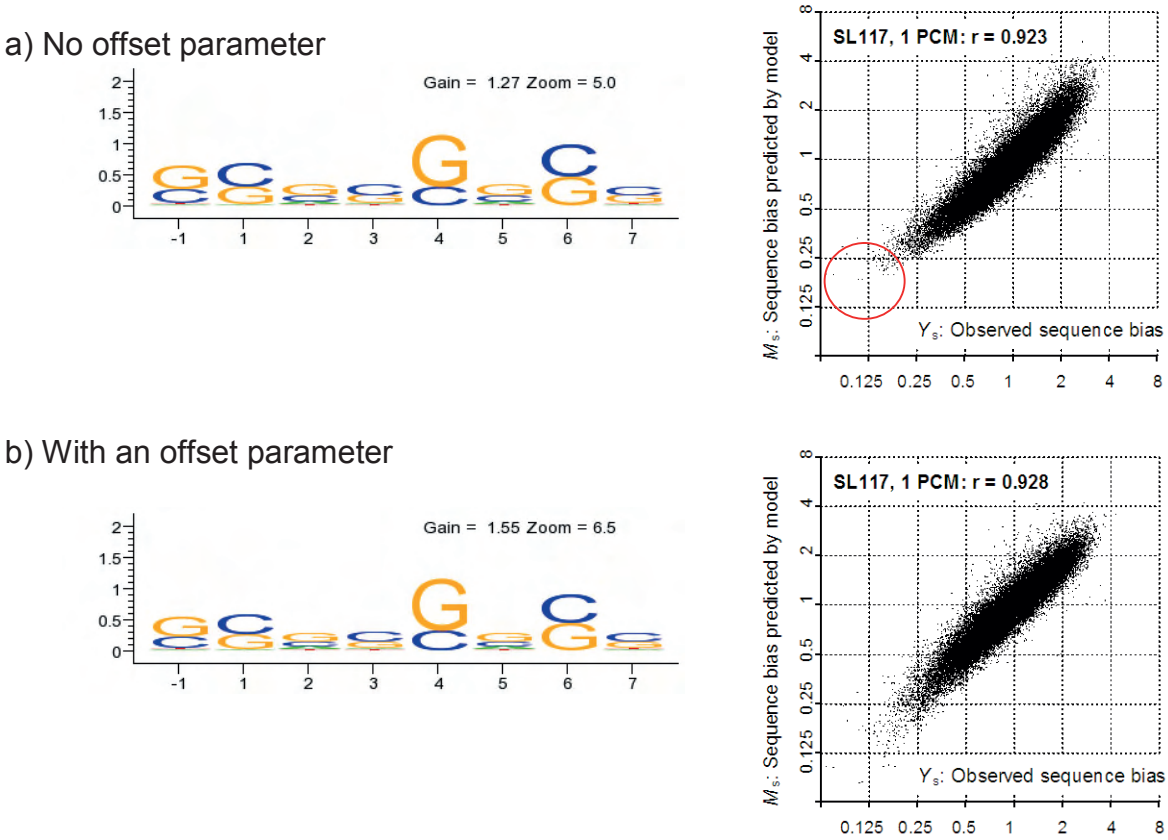


Figure 2-24 Model fitting of SM217 data improved through the use of an offset parameter. In each case the PCM generated as a result of model fitting is shown, together with a graph showing the correlation between model predictions and observed results. The red circle indicates the region that is most improved by the addition of the offset.

2.4.3 The problem of determining the optimal number of PCMs ‡

When creating models to fit observed data it is always necessary to consider the appropriate degree of model complexity given the data available and the problem being considered.

The general assumption is that the model should be as parsimonious as possible whilst still achieving a good fit between the model predictions and the observed data. The parsimony applies both to the complexity of the algorithm and also the number of free parameters within the algorithm.

One approach to choosing the optimum number of parameters, based on a Bayesian approach, gives rise to the Bayesian Information Criteria (BIC) which is a function of L, the likelihood of the result when the model parameters give the maximum model likelihood, k, the number of parameters and n, the number of data-points being fitted [89].

$$BIC = k\ln(n) - 2\ln(L) \tag{2.37}$$

If it assumed that there is a normal distribution of data points around that predicted by the model with a standard deviation of $\sigma_{\mathbf{s}}$ then the probability of the data given the model is given by:

$$P(data|model) = \prod_{\mathbf{s}} \frac{1}{\sigma_{\mathbf{s}}\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{y_{\mathbf{s}} - f(\mathbf{s})}{\sigma_{\mathbf{s}}}\right)^2\right] \tag{2.38}$$

The probability of the model given the data, required in order to determine the maximum likelihood is then

$$\begin{aligned} P(model|data) &= \frac{P(data|model)P(model)}{P(data)} \\ &= K \bullet P(data|model) \end{aligned} \tag{2.39}$$

from which the log likelihood is given by

$$\begin{aligned} \log\left(P(model|data)\right) &= \log(K) + \log\left(\prod_{\mathbf{s}} \frac{1}{\sigma_{\mathbf{s}}\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{y_{\mathbf{s}} - f(\mathbf{s})}{\sigma_{\mathbf{s}}}\right)^2\right]\right) \\ &= \log(K) + \log\left(\prod_{\mathbf{s}} \frac{1}{\sigma_{\mathbf{s}}\sqrt{2\pi}}\right) + \log\left(\prod_{\mathbf{s}} \exp\left[-\frac{1}{2}\left(\frac{y_{\mathbf{s}} - f(\mathbf{s})}{\sigma_{\mathbf{s}}}\right)^2\right]\right) \\ &= \log(K) + \sum_{\mathbf{s}} \log\left(\frac{1}{\sigma_{\mathbf{s}}\sqrt{2\pi}}\right) + \sum_{\mathbf{s}}\left[\log\left(\exp\left[-\frac{1}{2}\left(\frac{y_{\mathbf{s}} - f(\mathbf{s})}{\sigma_{\mathbf{s}}}\right)^2\right]\right)\right] \\ &= \log(K) + \sum_{\mathbf{s}} \log\left(\frac{1}{\sqrt{2\pi\sigma_{\mathbf{s}}^2}}\right) + \sum_{\mathbf{s}}\left[-\frac{1}{2}\left(\frac{y_{\mathbf{s}} - f(\mathbf{s})}{\sigma_{\mathbf{s}}}\right)^2\right] \end{aligned} \tag{2.40}$$

In these investigations the model fitting has been achieved by minimising a function of the form $\sum_{i=1}^{n}\left(y_i - f(x)\right)^2$ where $y_i = \log_2(M_s)$ and $f(x) = \log_2(Y_s)$. If it is assumed that the standard deviation of the distribution is the same for all of the sequences then this model fitting will determine the model with the maximum likelihood and which the parameter likelihood L can be shown to be given by

$$\log(L) = \log(K) + n\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2\sigma^2}\sum_{\mathbf{s}}\left[\left(y_{\mathbf{s}} - f(\mathbf{s})\right)^2\right] \tag{2.41}$$

From which, by combining with (2.27), the BIC can be written as

$$\begin{aligned} BIC &= k\ln(n) - 2\left(\log(K) + n\log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) - \frac{1}{2\sigma^2} \sum_{\mathbf{s}:N_{\mathbf{s}}>T}^{n} \left(\log_2(Y_{\mathbf{s}}) - \log_2(M_{\mathbf{s}}) \right)^2 \right) \\ &= k\ln(n) - 2\log(K) - 2n\log\frac{1}{\sqrt{2\pi\sigma^2}} + \frac{1}{\sigma^2} \sum_{\mathbf{s}:N_{\mathbf{s}}>T}^{n} \left(\log_2(Y_{\mathbf{s}}) - \log_2(M_{\mathbf{s}}) \right)^2 \\ &= k\ln(n) - 2\log(K) - n\log\frac{1}{2\pi\sigma^2} + \frac{1}{\sigma^2} \sum_{\mathbf{s}:N_{\mathbf{s}}>T}^{n} \left(\log_2(Y_{\mathbf{s}}) - \log_2(M_{\mathbf{s}}) \right)^2 \end{aligned} \tag{2.42}$$

The maximum likelihood estimate of the variance is given by

$$\sigma^2 = R/n \tag{2.43}$$

where R is the residual sum of the squares, i.e.

$$\sigma^2 = \frac{1}{n} \sum_{\mathbf{s}:N_{\mathbf{s}}>T}^{n} \left(\log_2(Y_{\mathbf{s}}) - \log_2(M_{\mathbf{s}}) \right)^2 \tag{2.44}$$

Substituting into (2.42) then gives

$$BIC = k\ln(n) - 2\log(K) - n\log\left(\frac{n}{2\pi \sum_{\mathbf{s}:N_{\mathbf{s}}>T}^{n} \left(\log_2(Y_{\mathbf{s}}) - \log_2(M_{\mathbf{s}}) \right)^2} \right) + n \tag{2.45}$$

The values of the BIC for the SL117 and SL523 datasets have been calculated for various numbers of PCMs after the model fitting has been completed to determine the optimal number of PCMs (2-25). The value of 2 log(K) has been ignored as this is a constant for a particular dataset and therefore makes no difference to the determination of the minimum value of the *BIC*. In this case each PCM consists of coefficients for eight nucleotides, and each nucleotide can be considered as being modelled by three parameters, in that the modelling includes the constraint that the weights for each of the four nucleotide types sum to unity. This leaves three free parameters per nucleotide. There is in addition an overall scaling factor k parameter for each of the PCMs, resulting in a total of 25 parameters per PCM. n is the dataset size of 37174. It is this size rather than 65536 because of the use of a threshold to exclude sequences that occur fewer than 50,000 times within the genome.

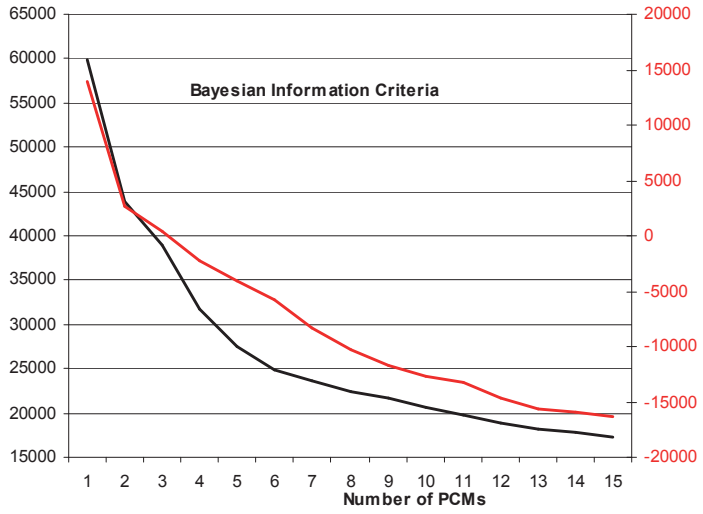


Figure 2-25 Variation of BIC and Pearson coefficient with the number of PCMs. The values for SL117 (black) and SL523 (red) datasets are shown with varying numbers of PCMs incorporated into the model.

SL117

	BIC	Pearson
1	13939	0.9202
2	2644	0.9423
3	432	0.9462
4	-2202	0.9504
5	-4019	0.9532
6	-5766	0.9557
7	-8349	0.9591
8	-10349	0.9616
9	-11710	0.9632
10	-12688	0.9645
11	-13270	0.9653
12	-14610	0.9668
13	-15610	0.9679
14	-15855	0.9684
15	-16323	0.9690

SL523

	BIC	Pearson
1	59956	0.8128
2	43924	0.8829
3	39022	0.8987
4	31660	0.9181
5	27464	0.9276
6	24910	0.9330
7	23599	0.9358
8	22396	0.9383
9	21682	0.9399
10	20562	0.9422
11	19816	0.9437
12	18810	0.9457
13	18216	0.9469
14	17901	0.9477
15	17285	0.9490

The figures and graphs show that the BIC values continue to decrease over the range of PCMs considered, indicating that the optimal number of PCMs is greater than 15. This can be explained partly because the dataset size n of 37174 is relatively large. If an additional PCM results in a relatively small average improvement in the fit of ε for all of the points then the $\sum_{i=1}^{n} \frac{\left(y_i - f\left(x\right)\right)^2}{\sigma_i^2}$ term decreases by n times ε which can still result in a significant change in the value of this term, even for relatively small values of ε. Consequently the effect of even a very modest improvement in fit associated with an additional PCM will continued to be more significant than the penalty of $k \log(n) = 263$ that is introduced as a result of adding k = 25 extra parameters.

One indication more parameters being used in the model than is appropriate in the model is the presence of overfitting, where the model fits to the random noise in the data. This can be checked for using cross validation, where the model fitting is performed with a subset

of the data and checked using a different subset. In this case the chromosomes were divided into two groups such that the total number of nucleotides in each group was approximately equal. Model fitting was performed using the sequence bias data from the first group, and then the Pearson Correlation Coefficient calculated for the fit between model and each of the two data groups.

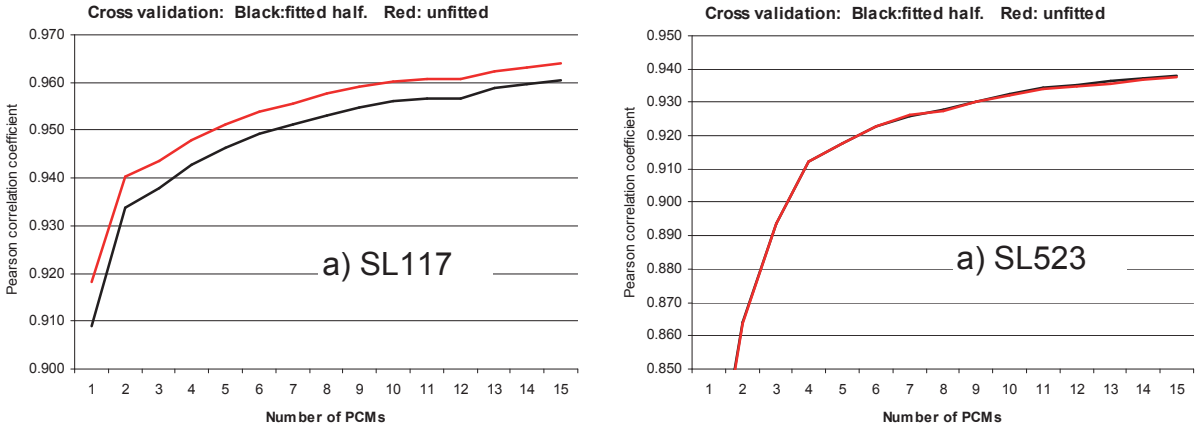


Figure 2-26 Cross validation shows no over fitting with up to 15 PCMs Pearson coefficient values for data from the 'half' genome used for model fitting (black) and the 'half' genome not used for modelling (red). The continuing improvement in fit for the half genome not used for model fitting indicates that there is no significant overfitting for these datasets when up to 15 PCMs are used in the model.

The results (Figure 2-26) show that there is no over fitting for the two reference datasets, SL117 and SL523, when up to 15 PCMs are used in the model. This is consistent with the BIC results which also indicate that the optimal number of PCMs for fitting the data is greater than 15.

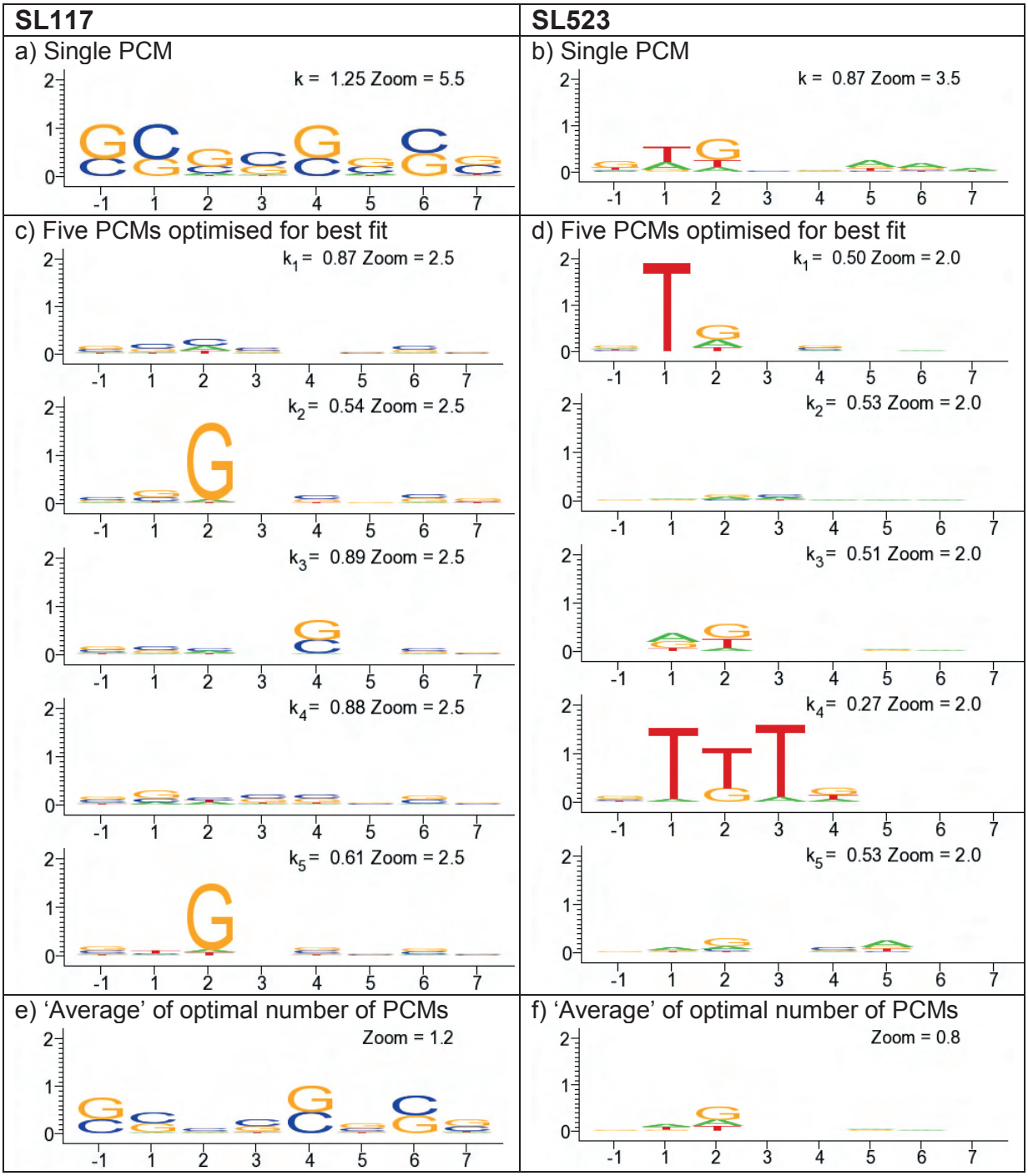


Figure 2-27 The effect of adding extra PCMs differs between experiments a & b) PCMs obtained when a single PCM is used for model fitting. c & d) Optimal set of five PCMs for the two sets of data. e & f) A single PCM generated as a vector sum of the five PCMs, when mapped to their 3D equivalents, giving an indication of the 'average' of the PCMs.

Figure 2-27 shows the results of fitting a single PCM, and five PCMs to the SL117 and SL523 data. This demonstrates that the model fitting algorithm takes advantage of the

additional PCMs in very different ways in the two datasets. In the case of SL523, the additional PCMs are consistent with the hypothesis that there are multiple alternative nucleotide sequences that give rise to an increased probability of DNA fracture, which are only identified when the model is given the capability of additional PCMs.

In the case of SL117, the additional PCMs do not suggest that there are radically different alternative sequence biases. Instead it suggests an inadequacy in the way the information associated with a PCM is used to calculate the fracture probability, which the model fitting algorithm overcomes by creating a set of variants of the single PCM with different distributions of weightings applied to the original single PCM.

This suggests that a modification to the algorithm for calculating fracture probabilities from a set of weights in a PCM could well achieve an equivalent improvement to the fit that was achieved by incorporating additional PCMs. As with the results of Section 2.4.2, a few simple modifications to the algorithm were investigated, but none were found that delivered a significant improvement in the model fit (data not shown).

In these examples, the weightings for a single PCM are used when the data exhibit a characteristic such as in SL117 as this is sufficient to indicate the character of the data. If the characteristics are similar to those shown by SL523 then sufficient PCMs to give an indication of the variation in PCM patterns are used.

2.4.4 Previous analyses of sequence bias by Schwartz et al missed key features ‡

Introduction

One of the few previous analyses of sequence bias in ChIP-seq data was by Schwartz et al. [88] which was part of a wider analysis of sequence bias in both ChIP-seq and RNA-seq data. The ChIP-seq data that they chose to analyse was one specific dataset that had been created as part of an earlier investigation of the genome wide investigation of HATs and HDACs by Wang et al. [99]. The paper does not state why this particular dataset was chosen. The following reanalysis shows that by picking this particular dataset and analysing it in the way that they did, they missed some of the significant features described elsewhere in this chapter.

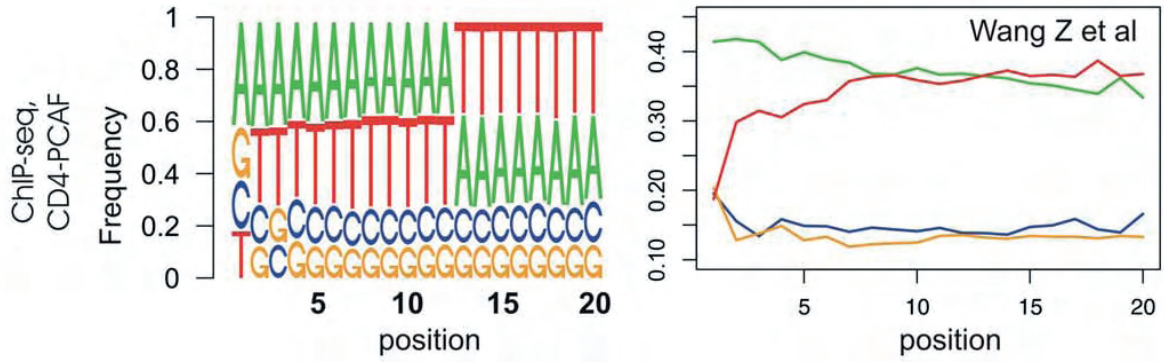


Figure 2-28 Extract from Figure 1 of Schwartz et al. The diagrams represent the sequence bias of the first 20 nucleotides of the fragments in two different ways, showing a bias towards As and Ts.

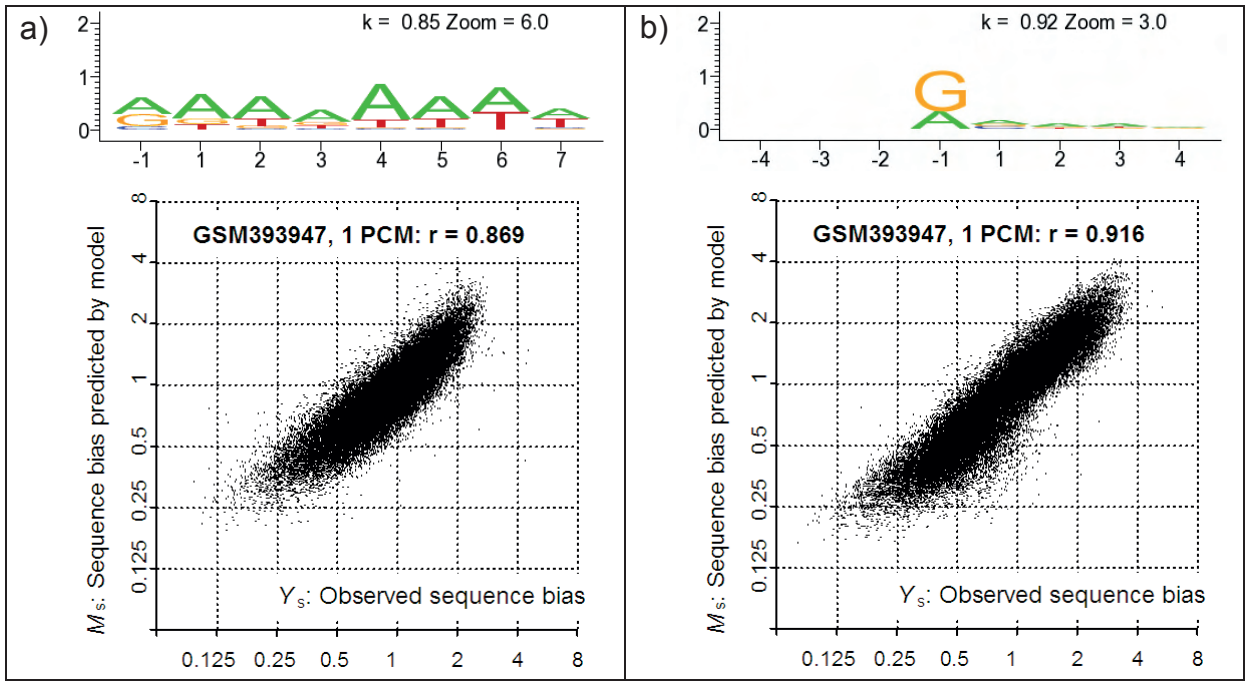


Figure 2-29 Sequence bias for GSM393947 shows additional features not identified by Schwartz et al. a) PCM that just includes the nucleotides at the start of the fragment. This corresponds to the start of the region examined by Schwartz el al. and is similar to the results they obtained. b) Bias of the nucleotides +/-4 nucleotides from the fragment start, showing a strong G bias in the nucleotide immediately before the start of the fragment, which was not seen in the previous analysis.

The dataset that was chosen by Schwartz et al. was created as a result of the sequencing of immunoprecipitated DNA fragments with bound PCAF protein from CD4 cells and was lodged in the GEO database as GSM393947. Figure 1 from the Schwartz paper shows the nucleotide bias they observed at the start of the fragments, showing a bias towards A at the

first nucleotide position of the fragment, and a strong bias to both A and T in the subsequent nucleotides (Figure 2-28).

Figure 2-29a shows the sequence bias for the first eight nucleotides of the fragment as determined using the methods outlined in this chapter. It matches the Schwartz data in showing a bias towards A and T at the start of the fragment, with the T bias being less significant in the first nucleotide. Figure 2-29b then shows the sequence bias for the 8-mer starting four nucleotides before the start of the fragment and a somewhat different picture emerges, in that it picks up a strong G bias in the first nucleotide before the start of the fragment.

Results: Analysis of GSM418301: HDAC binding control data HeLa cells

Given the unusual AT preference of the sequence bias in the SM393947 dataset, it was felt worth looking at some of the other data produced during the investigation by Wang et al. A second dataset, the control data for the examination of HDAC binding in HeLa cells, accession number GSM418301, was investigated (Figure 2-30). This shows the same tendency for a G immediately prior to the fragment start, but a very different characteristic within the fragment, with a pattern of multiple alternative biases. This is somewhat different from the dataset chosen by Schwartz et al. for their investigation of bias in fragmented DNA, and similar to datasets investigated elsewhere in the chapter.

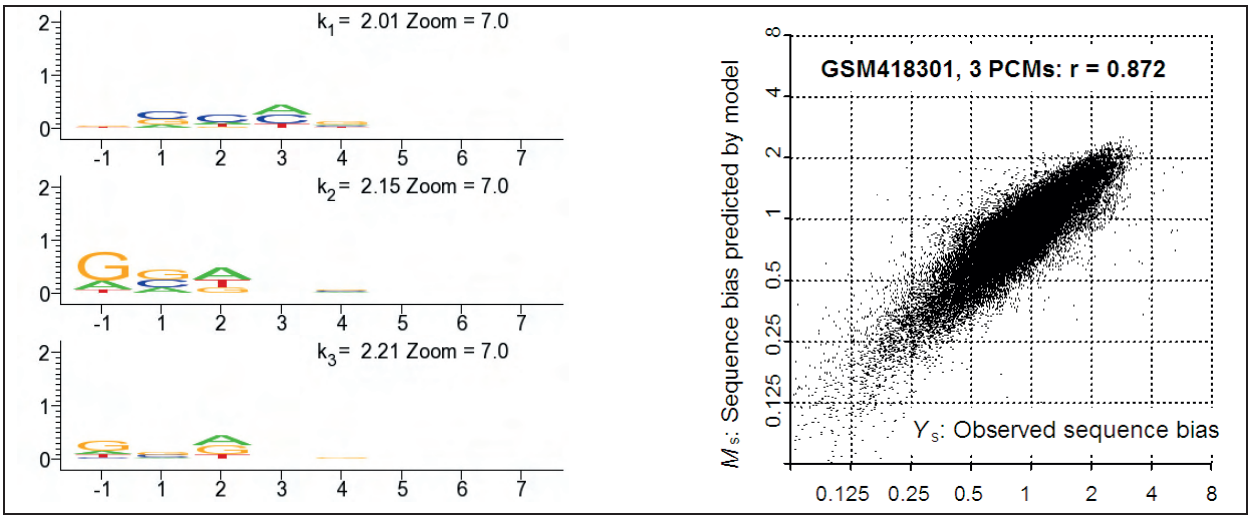


Figure 2-30 Analysis of region sequence bias in GSM418301. The results for this dataset are more consistent with results from other datasets than was the case for the GSM393947 data.

Conclusions

If this result is compared to the results of the other datasets analysed in this chapter it can be seen that the strong AT bias in the Schwartz et al. results is a very unusual

characteristic. However, is the G preference just before the start of the fragment is not inconsistent with a number of other sets of results (e.g. Figure 2-13c). This means they missed what now appears to be one of the common ChIP-seq bias characteristics as a result of their choice of method which did not account for any possible influence of nucleotides before the start of the fragment. An analysis of other datasets from the same source would have highlighted that there was greater sequence bias variation than appeared from the dataset analysed. Consequently, with hindsight, it was perhaps misleading for this paper to base a conclusion on this single result and in doing so the authors were apparently not aware of the significant variation that exists between different experiments.

2.5 Discussion †

It has previously been acknowledged that there is a bias in the nucleotide composition both throughout and at the start of sequenced fragments from procedures that involve fragmenting DNA [25, 26, 88]. Such biases will reflect biases in the location where the DNA originally fragments, together with biases introduced during the subsequent conversion amplification and sequencing of the fragments.

The modelling technique presented in this paper shows that for both RNA and DNA this bias is a mixture of multiple alternative patterns of nucleotide weightings.

2.5.1 ChIP-seq data show an unexpected asymmetric sequence bias around the fragment start position †

This analysis has thrown up two aspects to the nucleotide bias in ChIP-seq data that would not be expected from a simplistic model of what happens at the molecular level during the ChIP-seq process.

The first is the lack of symmetry of the sequence bias around the position of the start of the fragment that is often observed. A simple picture of the process would suggest that when the DNA cleaves during fragmentation, the DNA on either side of the fragmentation site is equally likely to become a fragment that is ultimately selected for sequencing. This picture would suggest that any nucleotide bias would therefore be symmetrical around the fragment start. However many of the ChIP-seq results, such as SL523, shows extreme asymmetry, with the nucleotide bias occurring predominantly on the side of cleavage site associated with the fragment that is ultimately sequenced.

2.5.2 ChIP-seq data show an unexpected variety of different sequence bias patterns

The second unexpected aspect is the variety of different biases exhibited for the different sets of data. Within the variety the data can be arranged into groups with common characteristics. For example the Yale lab data SL102 to 218 shows a specific GC-rich pattern (Figure 2-11b), whilst other data shows a different multiple bias characteristic (Figure 2-12a and b).

One possible explanation for the observed asymmetry is that bias is introduced by the processing of the fragments after cleavage. An investigation of such potential biases [2] identified a significant bias as a result of the suppression of GC-rich fragments during the PCR amplification stage. The degree of suppression was a function of the PCR cycling. In some conditions fragments with a GC content of >65% or <11% were suppressed by a factor of 100.

Such PCR induced bias would not appear to explain the results obtained here, in that some of the asymmetrical examples (Figure 2-11b) involve an increased abundance of fragments that are GC-rich at their fragment ends, rather than the suppression of such fragments. Furthermore, in the Myers lab data there were two different PCR protocols that were used but although different bias characteristics were seen, the variation in bias does not align with the usage of the two protocols.

Another possibility is that the asymmetry might be associated with specific cell lines or cell treatments. However, examples such as SL117 and SL523 suggest that this is not the case as they show very different bias characteristics even though they involve identical cell lines and treatments.

Another plausible hypothesis is that the asymmetrical bias is associated with the process of fragmentation. The Myers lab notes that the sonication process was changed in the fall of 2009 (Section 2.2.1) but does not record the change against the data. It is the case that the early results (including SL117) show the GC-rich fragment ends, and the later results (including SL523) show the more variable biases in the fragment ends, and this is not inconsistent with the change in nucleotide bias being associated with the change in the sonication process.

2.5.3 GC-rich bias arises from GC cleavage preference †

One characteristic identified by this analysis is that some samples show an increased likelihood for fragments starts to be in a locally GC-rich location, with the presence of GC

nucleotides up to two nucleotides away from the cleave site appearing to increase the probability of DNA fragmentation (Figure 2-12c and d). A recent observation that DNA tetranucleotides were more likely to cleave at a CG bond when exposed to ultrasound suggests that in some circumstances the same mechanism may be the dominant mechanism in determining the cleavage locations during ChIP-seq sonication [37].

2.5.4 GC-rich fragment ends may propagate through G-quadruplex formation †

Figure 2-13b) is an example of where a slight asymmetry is beginning to develop in this pattern, with additional bias appearing further into the sequenced fragment, this being on the way to the full asymmetrical GC bias that can be seen in Figure 2-11a and b).

If the observed asymmetry is associated with the DNA fragmentation stage of the process then one way in which such asymmetry might arise is if newly created fragment ends are able to catalyse the creation of further fragments that end with a similar nucleotide sequence.

Such a model would involve the ends of double stranded fragments aligning to a location elsewhere in the DNA where there is a degree of sequence similarity, forming a short quadruplex- like structure. At the end of the catalysing fragment there would be an abrupt transition to a conventional double stranded DNA structure and it is possible that some aspect of this transition increases the likelihood of the double stranded DNA fracturing at this location, creating a build-up of fragments with similar end sequences.

The existence and biological significance of DNA quadruplex structures consisting of four parallel/anti-parallel DNA strands is recognised [79], and it is possible that the specific constraints of such a structure are responsible for the very specific pattern of GC bias seen in SL117 and similar examples in the supplementary data.

A combination of the tendency for DNA to fracture at GC-rich locations, together with the preference for Guanine nucleotides in G-quadruplexes [79] could therefore be responsible for the build up of fragments with a very characteristic GC-rich pattern at their ends.

2.5.5 Propagation of non GC-rich fragment ends may also involve quadruplex formation †

However, such a model does not account for the build up of fragments with a range of different sequences at the fragment ends, as is seen in the later data from the Myers lab, and also the majority of the rest of the data analysed. The quadruplex model previously proposed would now require that the conditions in these experiments allow for the creation of

quadruplexes which were not so dependent on GC-rich sequences. This would allow any fragment sequences to catalyse the creation of further fragments whose sequences matched. Such a model could provide an explanation for the consistency in sequence bias within experiments (Figure 2-9), but the variability between experiments, in that the fragment end sequences seen would be very dependent on which sequences emerge in the random fragmentation that occurs at the start of the process.

Section 2.3.7 provides evidence for one way in which alternative over-represented sequences can arise in that the Y1109-1 data shows a clear tendency for fragments to start with the repeating motif TGGAA. This sequence is a major component of human centromeres [103], and in this experiment these regions could have provided the seeds for the over-representation of fragments starting with this motif which then originate not only from the centromere but from locations with similar sequences throughout the rest of the genome.

2.5.6 Input data are unsuitable for use as a reference for sequence bias compensation of data from immunoprecipitated fragments

The poor correlation between the sequence bias of input and immunoprecipitated data (Section 2.3.8) could be because the fragments associated with the two datasets were drawn from different pools with different sequence bias characteristics, or because of some effect associated with the process of immunoprecipitation. This makes this input data unsuitable for use as reference data for performing sequence bias compensation on the immunoprecipitated data.

There are other factors that weigh against the use of input data as the general approach for compensating for sequence bias.

An examination of the contents of many of the public repositories of ChIP-seq data such as the NCBI GEO database or the ENCODE data suggests that the input and control data do not come from the same experiment, in that there is frequently no clear link between the replicate identities for the two sets of data. This can be because here are clear differences in labelling between the two sets of data or there are different numbers of replicates with no clear association between the two sets of replicates. This frequently arises when a series of ChIP-seq experiments are performed for a particular cell line each with a different antibody to select fragments with different bound proteins. The implicit assumption is that a representative input dataset is satisfactory to be used for any of these. This makes it very difficult to determine which input data to use when using publically available datasets.

Secondly, where there does at first sight appear to be a suitable set of input data, a lack of correlation between sequence bias in input data and immunoprecipitated data in many publically available datasets still casts doubt on the appropriateness of using this data, as was the case for the SL522/523 sets of data. Finally, there is an increasing tendency for data to be published without a corresponding input or control dataset as is the case of the hg19 ChIP-seq data published as part of the ENCODE project.

2.5.7 Fragmentation in GC-rich sequence may be associated with CG dinucleotide underrepresentation in the genome

The complexity of the interactions that exist within the DNA in an organism creates the possibility for many indirect correlations between different characteristics, including the observed relationship between sequence bias and 8-mer frequency in the genome (Section 2.3.11). However the strength of the correlation makes it worth exploring how this may have arisen in case there is a direct connection between the two characteristics.

In the four datasets that have been examined, the correlation only exists when the ChIP-seq data has a sequence bias where fragment starts are more associated with a region that is richer in GC nucleotides . An implication of the results in Figure 2-20 is that these 8-mers with a higher GC content are under-represented in the genome, which is consistent with previous observations that the CG dinucleotide is under-represented in the genomes such as the human genome [36]. The underrepresentation of the CG dinucleotide is usually explained with the methylation-deamination-mutation hypothesis, where methylation of the cytosine in the dinucleotide CG to 5-methylcytosine by a methylase and subsequent deamination to thymine results in the conversion of a CG dinucleotide to a TG/CA dinucleotide. Support for this hypothesis comes from the fact that *Drosophilia* and *C. elegans* do not possess a methylase enzyme and also do not show a CG underrepresentation. While the methylase mechanism may play a role in reducing the numbers of CG dinucleotides, it has also previously been noted that it is not a completely satisfactory explanation for CG underrepresentation in that CG dinucleotides are also underrepresented in mitochondria, where there is no methylase activity [36].

The result in 2.3.11 may suggest another mechanism that might contribute to creating the non-uniform distribution of dinucleotides that is seen in some species. It has been suggested that the sequence bias observed in some ChIP-seq experiments arises because the DNA is more likely to break at a CG rich dinucleotide (Section 2.5.3), which is consistent

with other results which show that this dinucleotide is vulnerable to fragmentation when exposed to ultrasound [37].

It is possible that the mechanism that underlies the cleavage susceptibility that exists during sonication is also active within the DNA in the cell, making short sections of Cs and Gs more vulnerable to breakages and mutations such that over evolutionary timescales these will come to be under-represented within the genome.

In the small number of samples examined, the data from *C. elegans* did not show the same relationship between the numbers of N-mers in the genome and the sequences bias as was shown with the *H. sapiens* data (Figure 2-21) and also did not show sequence bias towards fragment starts in CG rich regions (Figure C-8). While this is consistent with *C. elegans* not showing the an under-representation of the CG dinucleotide in its genome, significantly more data would need to be examined in order to see significance, of these small number of results.

2.5.8 Fragmentation in GC-rich sequences provide a possible explanation for poor quality *Arabidopsis* ChIP-seq data

Figure 2-31 shows that regions of high fragment density in the *Arabidopsis* input DNA often coincides with the location of exons, which can result in problems using the ChIP-seq protocol to locate protein binding regions away from exons where the fragment density is low.

While it is generally true that exons have a greater GC content than introns, this has been shown to be more pronounced in *Arabidopsis* [104]. Figure 2-31b) shows how fragment starts in *Arabidopsis* tend to be biased to locations with a high local GC content, which explains why the fragment density distribution predicted by the model aligns with the GC content in Figure c) and d). While this goes some way towards explaining the experimentally observed distribution, Figure 2-31a indicates that this is not a complete explanation for the distribution in that, for example, it does not explain why there are almost no fragments in some intergenic regions.

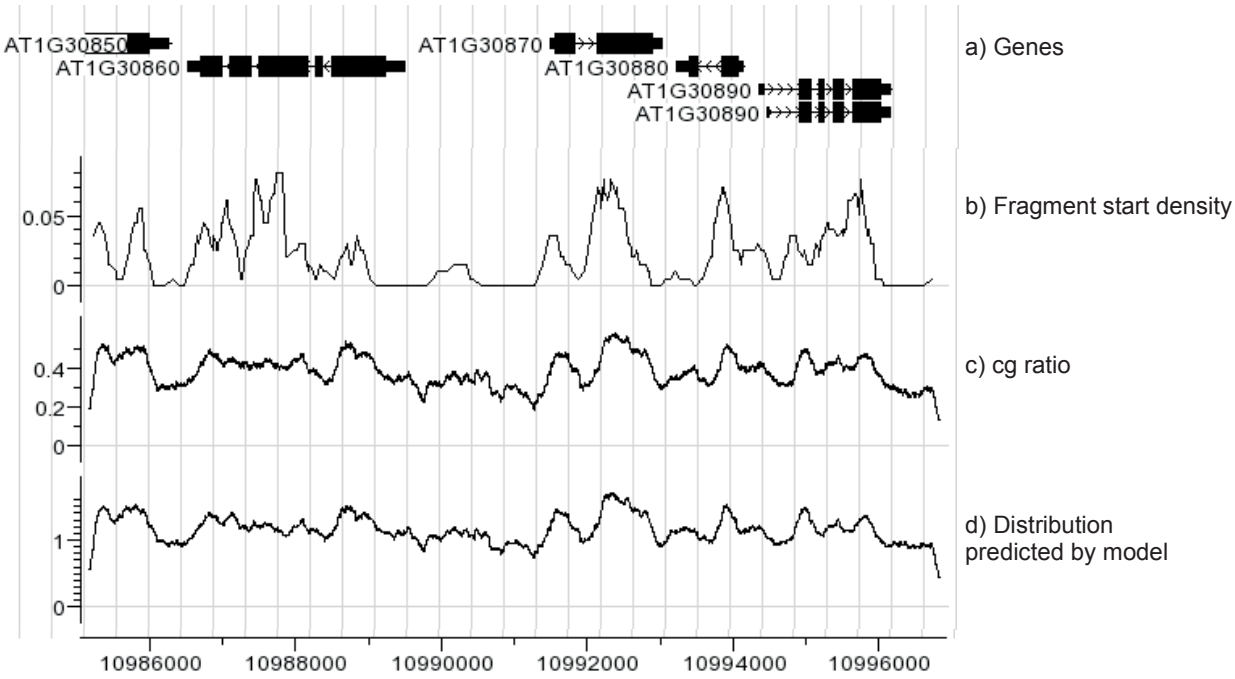


Figure 2-31 ChIP-seq fragment distribution in Arabidopsis. A region of chromosome 1 showing the gene distribution, Fragment start density of the input DNA, cg content and the fragment distribution predicted by the model using PCMs from Figure C-9)

A more even fragment distribution would improve the quality of binding site predictions in the regions where the fragment density is currently low. If the mechanism that causes the fragment distribution in the *Arabidopsis* experiments examined is linked to the tendency for fragment starts to occur in GC-rich regions then a better understanding of the mechanism that causes the different sequence bias characteristics may allow the fragment distribution to be changed to be a more uniform. Such a change would improve the quality of the results obtained from such experiments.

Chapter 3

Sequence bias in RNA-seq experiments

This chapter builds on the model-fitting approach discussed in the previous chapter and applies a similar technique to the analysis of the sequence bias at the start of RNA fragments from RNA-seq experiments.

3.1 Introduction

Section 1.5 has already provided a brief introduction to RNA-seq, with some coverage of the differences and similarities between this process and the ChIP-seq protocol. There has already been a number of publications relating to the sequence bias at the start of the sequenced fragments that has been found to exist in RNA-seq data [16, 40, 60, 88]. These publications showed that the bias has some of the characteristics of the bias in ChIP-seq data that was discussed in the previous chapter. In the case of RNA-seq data the primary mechanism previously identified as causing the bias is the selective binding of the random hexamers during the process of reverse transcription from RNA to DNA (Section 1-7)[40]. This has led to the development of a new flow cell reverse transcriptase sequencing (FRT-seq) protocol that has been demonstrated to reduce this bias quite significantly [68].

As well as investigating the bias in order to improve the protocols, the knowledge gained about the bias has been used to develop ways of manipulating the data to compensate for the effect of the bias [16, 40, 60, 88].

In all the previous work on RNA-seq data, the underlying assumption has been that there is a single nucleotide sequence pattern that describes the bias observed.

Here we demonstrate that, as with DNA fragmentation in the ChIP-seq protocol, the sequence bias characteristics that arise during RNA fragmentation and amplification are more complex than can be modelled using a single sequence pattern, and that modelling based on multiple alternative patterns within a single experiment provides a considerably richer and more detailed picture of what happens during the experimental procedure. It is then possible to distinguish between different sources of bias within the same experiment, and better understand the differences in bias between experiments that have been observed [88]. This work shows that in previous analyses, the effect of bias has sometimes been hidden as a result of using simple models that cannot capture the full complexity of the bias seen in RNA-seq experiments.

3.2 Method

3.2.1 Data sources ‡

This study used two sets of *Mus musculus* data from the Wold lab [73], one set of *Homo sapiens* data produced as part of an investigation into NF-κB binding [50] and two sets of data from *Arabidopsis thaliana* produced at the University of Warwick. In addition some published FRT-seq data has been analysed in order to contrast the bias that is seen with this new protocol against the bias observed when the traditional protocol is used. The details of the datasets used are as follows:

***Mus musculus* from the Wold lab**

These data were originally generated by the Wold lab [73] and submitted to the GEO Sequence Read Archive (SRA) database under study SRP000198, submission SRA001030; and subsequently used as part of an investigation into bias in RNA-seq data [60].

Name	SRX	SRA	Source
Mouse skeletal	SRX000352	SRR001361 & SRR001362	Skeletal
Mouse brain	SRX001866	SRR189589 & SRR006489	Brain

SRR189589 replaces SRR006488 which was originally submitted to the database and then subsequently withdrawn.

The sequence tags were aligned to the mm9 reference mRNA sequences created as part of the ENCODE project and downloaded from:

http://hgdownload.cse.ucsc.edu/goldenPath/mm9/bigZips/refMrna.fa.gz

Homo sapiens GSM484895

This is one of the datasets from the investigation into the variation of Pol II and NF-κB binding between 10 different individuals [50]. This dataset is part of the GEO dataset submission GSE19466. The sequences were aligned using Bowtie to the release 18 of the *Homo sapiens* reference mRNA sequences from the ENCODE database from

ftp://hgdownload.cse.ucsc.edu/goldenPath/hg18/bigZips/refMrna.fa.gz

GSM	filename
GSM484895	GSM484895_GM12878_RNAseq_rep1_FC42B8R_090629_s_7_eland_multi.txt.gz

Arabidopsis thaliana mRNA

These data are two technical replicates from an experiment to investigate the effect of *Botrytis* infection on *Arabidopsis* which were carried out at the University of Warwick. In both cases the cDNA was run on an agarose gel to select a limited range of fragment sizes. In the first replicate the selected band was centred on a fragment length of 200 bp. In the second replicate a band centred on 300bp was selected.

The sequence tags were aligned to TAIR10 representative cDNA sequences and the tags that did not align then aligned to the full cDNA sequence data. Both were obtained from:

ftp://ftp.arabidopsis.org/home/tair/Sequences/blast_datasets/TAIR10_blastsets

at www.arabidopsis.org

Name	GSM	File	Replicate
Arabidopsis 24hr Rep 1	GSM850477	SRR391051.sra	1
Arabidopsis 24hr Rep 2	GSM850478	SRR391052.sra	2

***Homo sapiens* ERA00183 using the FRT protocol**

This is one of the datasets submitted to the NCBI GEO short read archive as part of the data to support the FRT-seq protocol [68]. The sequences were aligned to the release 18 of the *Homo sapiens* reference mRNA sequences from the ENCODE database using Bowtie.

Submission	**Study**	**Run**	
ERA000183	ERX002245	ERR007690	ERR007690_1.fastq.gz ERR007690_2.fastq.gz

3.2.2 Fragment alignment

In each case the raw sequence data was aligned to cDNA sequence data in order to identify the distribution of fragments within the mRNA. This ignores the small proportion of mRNA that might originate from introns in unspliced RNA. As with ChIP-seq data, any fragments that align to multiple locations were ignored.

cDNA sequence data is available which provides the sequences for a number of transcript variants for each gene. It is also available in a form containing only the sequence of a single representative transcript variant for each gene. Aligning to the first sequence data results in fragments being discarded if they match to the orthologous sequences in alternative transcript variants. Aligning to the representative sequences results in fragments being discarded if they only align to a transcript variant other than the variant chosen as the representative transcript. In order to make best use of the data available, the sequence tags

were aligned to one of the cDNA sequence sets, and those fragments that did not align were then aligned to the other, and analysis performed with combined set of alignment data.

3.2.3 Analysis of RNA fragment start sites †

As with the analysis of ChIP-seq fragmentation, the underlying relationship that was modelled for RNA-seq data was $p(F|\mathbf{s})$, the probability of a fragment F being created given a specific genomic sequence **s** rather than $p(s|F)$, the probability of specific sequences being found at the start of a fragment (see Section 2.1.1).

The radically different characteristics of RNA-seq data required some significant changes to the analysis compared to that used for ChIP-seq data, whilst retaining the same approach of using one or more PCMs to model nucleotide bias in the region of the fragment start site.

Initial investigations suggested that $X_{g,x}$, the probability of a fragment starting at a specific location x in mRNA g could be represented by

$$X_{g,x} = G_g A_{g,x} Y_{\mathbf{s}(g,x)} \tag{3.1}$$

G_g is a measure of the expression level of the mRNA. $A_{g,x}$ is a function of x which varies relatively slowly along the mRNA sequence with a characteristic length of the order of 10 nucleotides, i.e. there is significant autocorrelation for values from positions that are less than 10 nucleotides apart, but the autocorrelation drops of for separations of greater then 10 nucleotides. The value of 10 was determined by visual inspection of the general characteristics of distribution of breaks within a gene. $Y_{\mathbf{s}(g,z)}$ is the same bias that was used previously as the weighting function for DNA sequences and is a function of the sequence **s** at position x in the mRNA sequence of gene g.

$A_{g,x}$ is not the subject of this investigation. The value of $G_g A_{g,x}$, the background fragment density incorporating both the expression level of the gene and the more slowly varying component of the fragment density, was therefore approximated from the observed data to generate a local background distribution in order that the function $Y_{\mathbf{s}}$ could be investigated.

The approximation used for $G_g A_{g,x}$, the background fragment density, is given by:

$$G_g A_{g,x} = {}^{1}\!/_{w^2} \sum_{i=-w}^{w} F_{g,x+i} \times (w - |i|) \tag{3.2}$$

$F_{g,i}$ is number of fragments starts at position i of gene g. This function creates a smoothed version of the fragment start distribution using a triangularly weighted filter of width 2w.

The same modified Amoeba optimisation algorithm [34] used for the ChIP-seq data was used to fit the parameters in order to minimise the error function E_{RNA} , defined as

$$E_{RNA} = \sum_{g=g_1}^{g_n} \sum_{x=1}^{L_g} \left(\begin{array}{l} \log_2 \left(G_g A_{g,x} M_{\mathbf{s}(g,x)} + P \right) - \\ \log_2 \left(X_{g,x} + P \right) \end{array} \right)^2 \tag{3.3}$$

g_1 and g_n are the first and last genes whose fragments are used to optimise the PCMs and L_g is the effective length of the associated reference mRNA. $M_{\mathbf{s}(g,x)}$ is the value produced by the model given the sequences **s** at position x in gene g, which is generated based on the coefficients of the PCMs.

The parameter P is a fixed offset that is added to reduce the effect of the noise associated with the genes or regions of the mRNA where there are small numbers of associated fragments.

3.3 Results

3.3.1 Modelling RNA fragmentation identifies regions with different bias characteristics ‡

RNA-seq data from *Arabidopsis thaliana* were used to study the relationship between the RNA fragment start positions and the RNA sequence.

Initial investigations suggested that there is a bias towards certain nucleotides up to 12 positions away from the start of the sequenced fragment from RNA-seq experiments. Model fitting was used to determine the PCMs required in order to achieve a reasonable fit between the observed data and fragment start distribution provided by the model.

The PCMs were optimised using the data from the eight genes for which the greatest number of mRNA fragments had been aligned. The Pearson coefficient was used to obtain a measure of the model fit for each additional PCM was added.

The use of the $G_g A_{g,x}$ multiplier within the equation (3.1) results in a degree of fit between the model and the observed fragment density even when $M_{\mathbf{s}}$ is set to a single scalar which does not provide any sequence dependency (Figure 3-1). The Pearson correlation coefficient in this case is 0.7018 when calculated for all the nucleotide positions in the top eight genes. This improves to 0.8878 when a single PCM is introduced into the model to give

a degree of sequence dependence. The improvement is clearly visible when plotted, showing the significant contribution made to the model by the introduction of per-nucleotide sequence dependency.

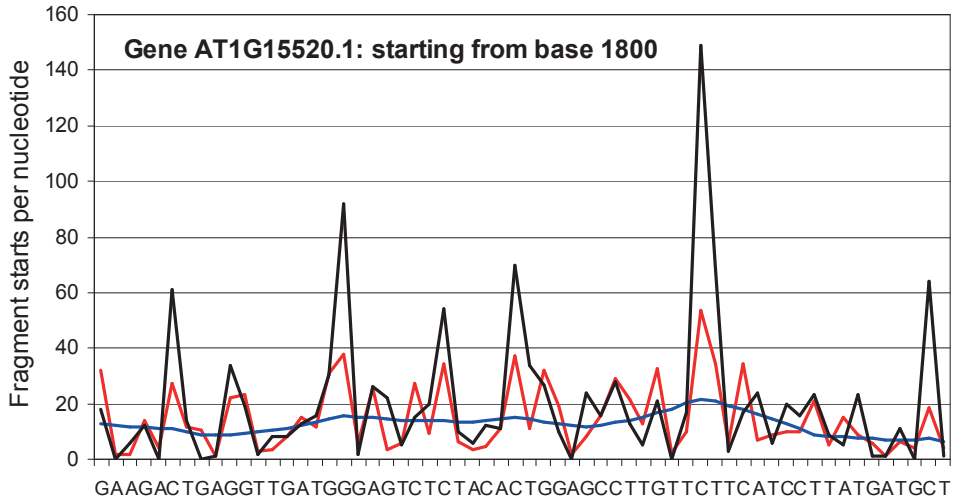


Figure 3-1 The output of a model with no sequence dependency still shows a small degree of correlation with the experimental data. Setting M_s to a single scalar (blue) shows some correlation with the observed fragment start density (black) although the correlation is considerably improved with the addition of the sequence dependence provided by a single PCM (red). Results obtained using *Arabadopsis* 24 hr rep 1 data

As well as the Pearson correlation coefficient being calculated for the first eight genes, whose data was used for the model fitting, the coefficient was also derived for the next eight genes to provide an indication as to whether the match for the first eight genes had been obtained as a result of over fitting. Over fitting would be indicated by a degraded fit for data from genes which were not used for the model fitting.

Note that the Pearson coefficients obtained from analysing the RNA-seq data are not directly comparable with those derived from ChIP-seq data in the previous chapter. This is partly because of differences in the way that model is being compared with the data. It is also the case that the correlation observed in the RNA-seq data is partly due to the use of $G_gA_{g,x}$, the background fragment density, as the input data to the model. This is derived from the experimental data and so is already correlated with this data. The Pearson coefficient is used to give an indication of the improvement to the match between model and experimental data generated by the model over and above that obtained using the $G_gA_{g,x}$, background

distribution, alone. This aspect of the modelling is specific to the RNA-seq data which is an additional reason why Pearson coefficient from the two types of data are not comparable.

These results suggest that the nucleotides are in two distinct regions with different characteristics (Figure 3-2). In the first region, extending for the first six nucleotides from the start of the fragment, the model required multiple PCMs in order to obtain a good match between the model and the observed data. The region from position seven onwards was different in character, in that all of the PCMs generated were almost identical, with a dominant U at position seven and an A at position 10, suggesting that a single PCM would be sufficient to define the character of this region.

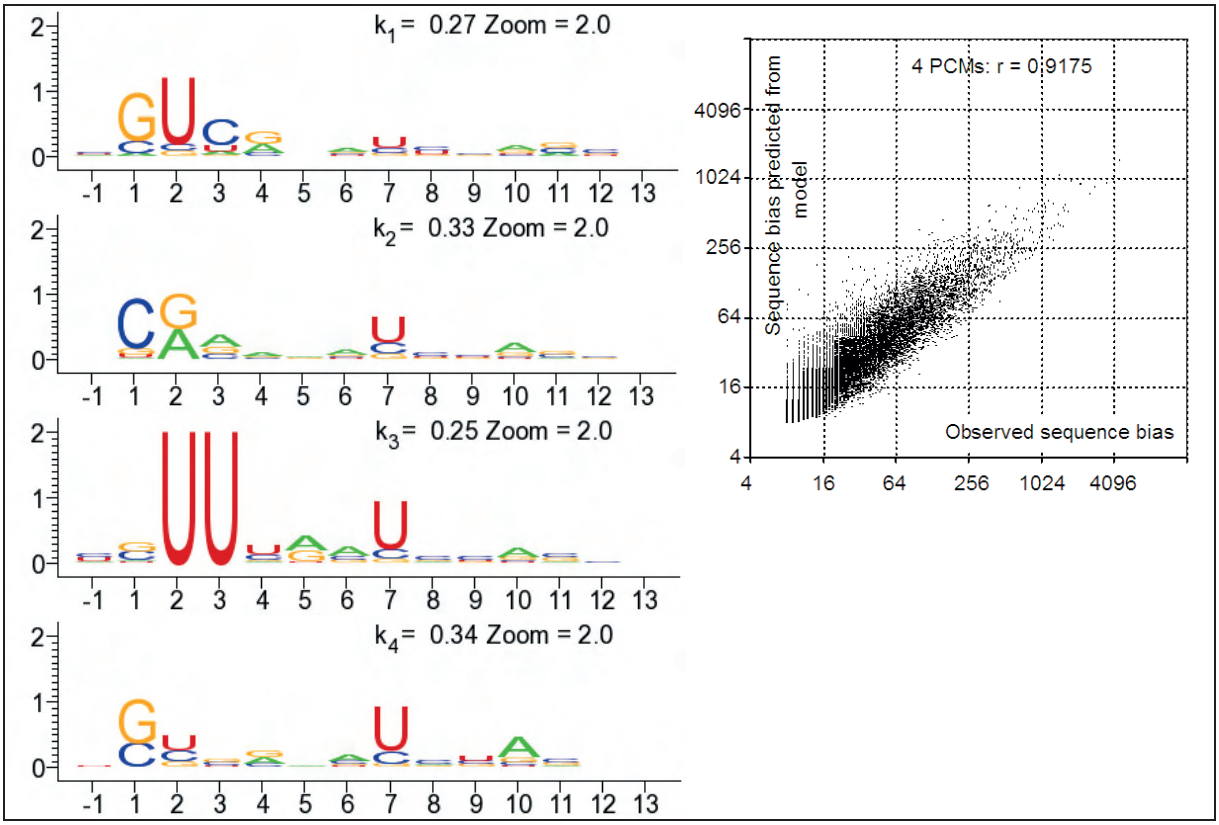


Figure 3-2 Four PCMs showing sequence bias for the first 14 nucleotides of RNA-seq fragments. PCMs cover the nucleotide before the fragment start and the first 13 nucleotides of the fragment. There is considerable variation associated with nucleotides one to six, and considerable similarity from seven onwards.

Consequently the model was changed so that there were multiple PCMs covering the region up until the 6th nucleotide, and the result from this component of the model is multiplied by a single PCM calculated using nucleotides 7 to 13, with PCM coefficients χ'_{i,n_i}.

$$M_{\mathbf{s}} = \max\left(k_j \prod_{i=-1}^{6} 4\chi_{i,n_i,j}\right)_{j=1}^{P} \times \prod_{i=7}^{13} 4\chi'_{i,n_i} \qquad (3.4)$$

The PCMs obtained with the new model reproduce the same variation in sequence at the start of the fragment that was seen previously, supporting the decision to only use a single PCM for the region from nucleotide seven onwards (Table 3-1)

	a) Multiple PCMs: positions -1 to 13		b) Multiple PCMs: positions -1 to 6 Single PCM: positions 7-13	
Number of PCMs	Genes 1-8	Genes 9-16	Genes 1-8	Genes 9-16
1 PCM	0.8878	0.8911	0.8898	0.8948
2 PCMs	0.9073	0.9084	0.9118	0.9141
3 PCMs	0.9143	0.9122	0.9190	0.9192
4 PCMs	0.9175	0.9136	0.9242	0.9246

Table 3-1 Pearson correlation coefficient indicating the fit between model and data for two different models. a) All PCMs independently cover the first 13 nucleotide positions. b) All PCMs only cover the first six nucleotides, there being an independent PCM covering position seven onwards. Model fitted using data from genes one to eight. Results for genes 9-16 verify that over fitting has not occurred, in that fitting improvements continue to be seen for these genes even though they did not contribute to the model fitting

The PCMs created by the model fitting process show multiple alternative bias patterns in the first six nucleotides (Figure 3-3), and the single pattern from nucleotide seven onwards matches the pattern seen in this region when multiple PCMs were used for this region.

.

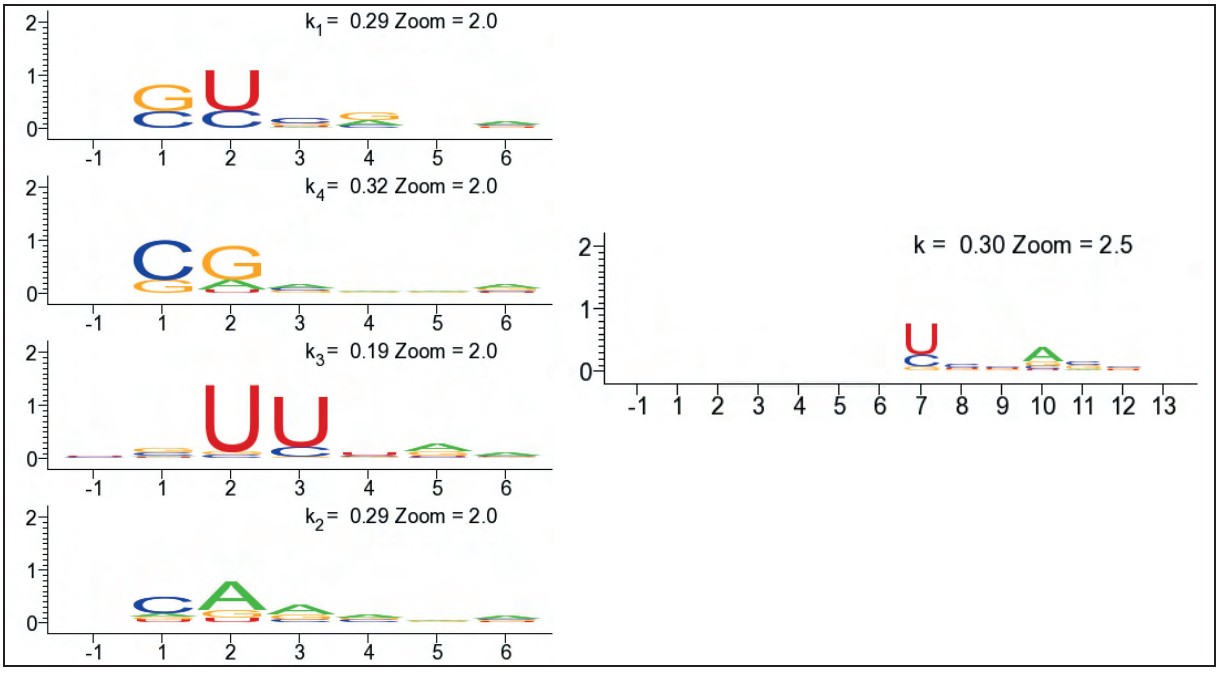


Figure 3-3 Four PCMs generated to match the first six nucleotides at the start of RNA-seq fragments, and a single PCMs cover the nucleotides from position seven onwards. There is a clear similarity between the four PCMs and nucleotides one to six of those shown in Figure 3-2.

A close match between the model and the observed data was achieved using data from the eight most highly expressed genes and a comparison with regions not used for model-fitting qualitatively demonstrates that the fit has not been achieved as a result of over fitting (Figure 3-4).

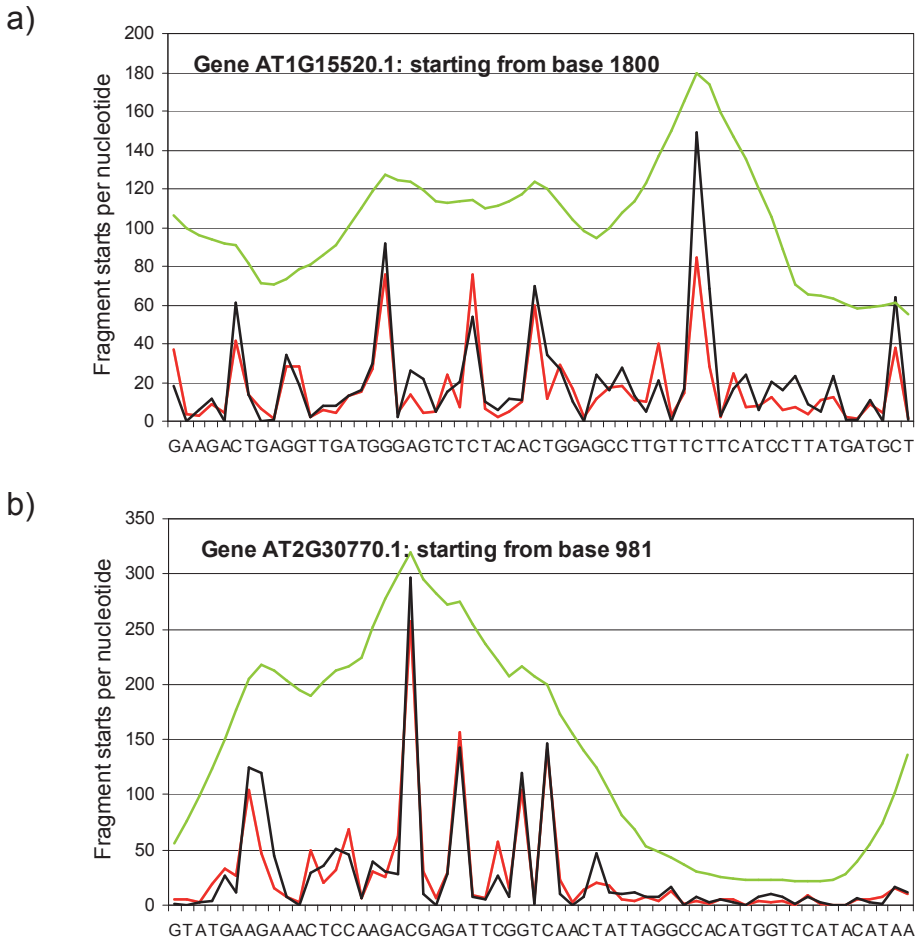


Figure 3-4 No evidence of over-fitting in regions not used for model fitting. a) A 50 nucleotide region of RNA-seq data showing a very good fit achieved between the observed fragment start density (black) and the value predicted from modelling (red). The value of the background density which is derived from the observed data and is the input fragment distribution for the model is shown in green b) Correlation of model prediction and observed data for a region that was not used for the original model fitting. All results obtained using *Arabadopsis* 24 hr rep 1 data.

When analysing potential over fitting in this way, one factor is the degree to which there are sequences in common between the data used for fitting and the data used for validation. If the full 14 nucleotide length sequence is considered then there are only 16 sequences in common between the sequences used for fitting and cross validation.

If the two 7-nucleotide sequences are considered independently then there are only 16384 permutations of a 7-nucleotide sequence, and the fitted and validation sequences are

approximately 15000 and 14000 nucleotides long respectively so the majority of sequences will appear in both datasets. However the fragment start statistics associated with these two sequence sets are still independent, making the second set of genes appropriate for use in cross validating the data.

3.3.2 The PCMs for the 5' and 3' ends of RNA-seq fragments are very similar

The results for the 3' ends of the RNA-seq data fragments from this experiment as well as results for *H. sapiens* and *Mus Musculus* RNA-seq data show that in any given experiment the reverse complement of the sequence bias at the 3' end of the fragment is remarkably similar to the sequence bias seen at the 5' end (Appendix D). The results also show that PCMs are always similar to those in figure 3-3, although there is variation in the detailed character of the PCMs for the first six nucleotides.

3.3.3 No over-fitting seen with RNA-seq data using up to nine PCMs ‡

Figure 3-5 shows the improvement in model fit for each additional PCM that is added to model the bias in the first six nucleotides. These results were obtained using 5' end sequences from SRX000352 (Mouse skeletal data: Wold lab).

The Pearson correlation coefficient and the E_{RNA} values have both been used to give an indication of the degree of fit achieved by the model. The model fitting was done with the eight genes with the highest RNA fragment density, and there is a clear improvement in fit for each additional PCM.

The Pearson correlation coefficient and E_{RNA} was also calculated for the next eight genes in order of fragment density. The data associated with these genes were not used for model fitting. The continuing improvement in the genes not used for model fitting indicates that the improvement with each additional PCM was not as a result of over fitting.

Figure 3-6 shows the PCMs obtained at the end of the process, showing that there is a predominant tendency for Cs or Gs at the start of the fragment, but that there is a continuously varying spectrum of nucleotide biases for the following nucleotides. The first two PCMs show a preference for Us in nucleotides two and three. Subsequent PCMs show a slight tendency for Cs in these positions and as this becomes more significant at the expense of a tendency for there to be a U, there emerges a tendency for them to be a G. There is always a tendency for there to be As at the end of the sequence, but this is slightly more significant when Gs dominate at the start of the sequence.

	Genes 1-8		Genes 9-16	
Number of PCMs	**Pearson**	**E_{RNA} error**	**Pearson**	**E_{RNA} error**
2	0.9285	7289.6	0.9213	4889.6
3	0.9340	6739.4	0.9283	4471.8
4	0.9379	6342.1	0.9307	4329.8
5	0.9387	6259.8	0.9311	4308.0
6	0.9399	6141.7	0.9315	4283.9
7	0.9409	6044.6	0.9327	4208.3
8	0.9422	5907.4	0.9345	4105.6
9	0.9427	5859.9	0.9346	4096.1

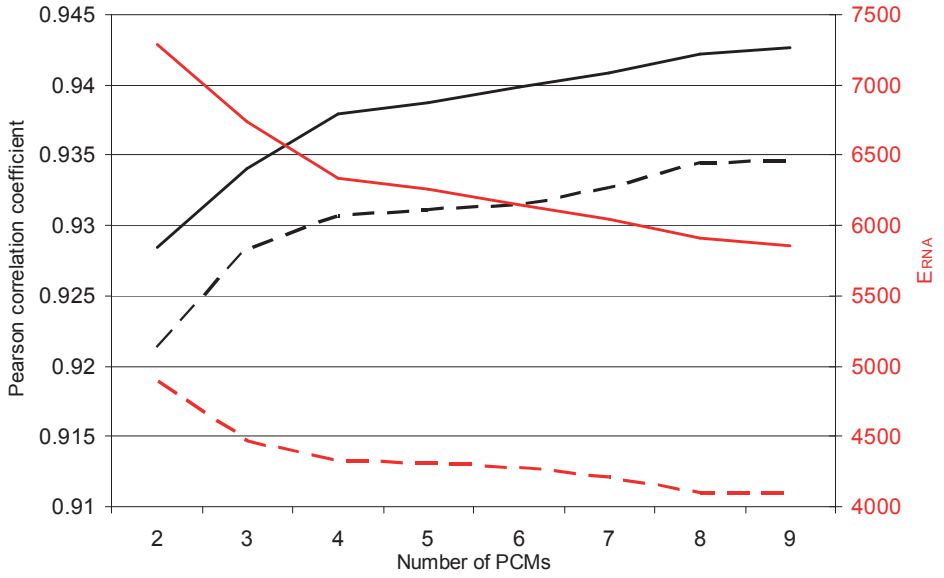


Figure 3-5 No over-fitting seen for up to nine PCMs. An increase in Pearson coefficient value and reduction in the E_{RNA} error value is seen with each additional PCM, both when calculated using the genes used for model fitting (1-8) and genes not used for model fitting (9-16). Genes numbered in decreasing order of the density of the fragments mapped to the gene.

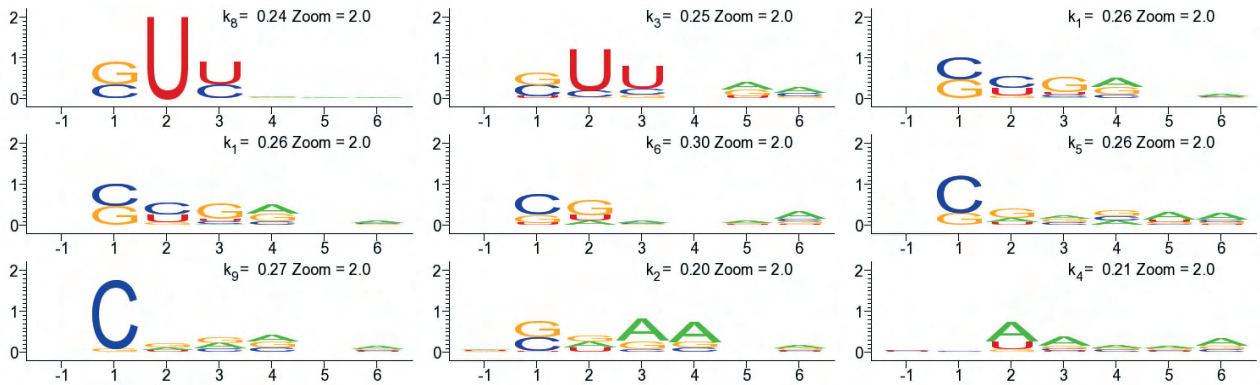


Figure 3-6 Nine PCMs obtained from model-fitting SRX000352 RNA-seq data. These give best match between model an observed data for the first six nucleotides.

It is possible to demonstrate the consistency between the picture of sequence bias that is seen with multiple PCMs and the simpler picture that emerges when only a single PCM is used for modelling. If the 3D vectors for the 9 PCMs are added together, giving an indication of the amalgamated effect of the PCMs, the resulting single PCM is very similar to that obtained if the modelling is performed using a single PCM (Figure 3-7).

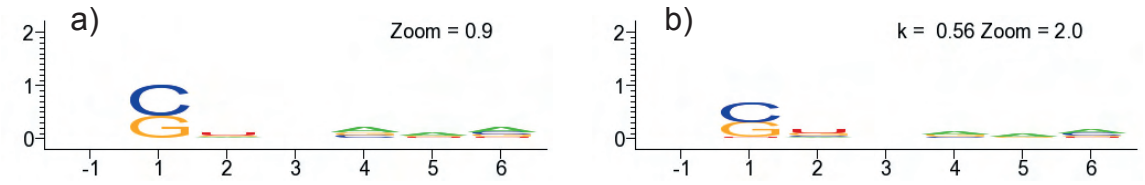


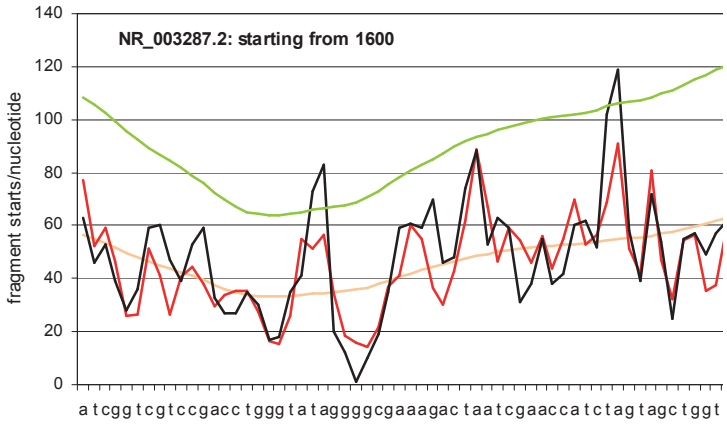
Figure 3-7 The results of single PCM model-fitting and from creating an 'average' of multiple PCM model fitting are very similar. a) Composite PCM constructed from the weighted vector sum of the nine individual PCMs, b) PCM created when only a single PCM is available for model fitting.

3.3.4 RNA-seq data processed using the FRT-seq protocol ‡

It has been proposed that the bias seen in RNA-seq data is as a result of a selectivity in the way the random primers bind to the RNA prior to the conversion to DNA [40]. In 2010, FRT-seq, a new protocol [68] for sequencing RNA-seq data, was introduced which was designed to overcome the bias that occurred in the previous protocol as a result of using random hexamer priming. In FRT-seq an adaptor sequence is ligated onto the 5' and 3' ends of the RNA fragments and the fragments are then attached to the flow cell where the fragment amplification proceeds. In the first stage of amplification reverse transcriptase is used to convert the RNA fragment to a DNA, and following stages of the process amplify the resulting DNA fragment.

Figure 3-8 compares the observed and modelled fragment distribution for the 5' ends of the fragments sequenced in dataset ERR007690, of submission ERA000183 to the European Nucleotide Archive (http://www.ebi.ac.uk/embl). This shows that although there is considerably less bias with the FRT-seq data, resulting in a much more even distribution of fragments, there is nevertheless some sequence dependent bias, and this can still be incorporated into the model to improve the model's ability to match the observed data.

a)



b)

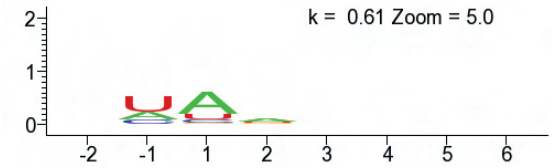


Figure 3-8 Model fitting of FRT-seq data 5' fragment end. a) Comparison between observed (black) and modelled (red) 5' fragment end distribution for a sample of FRT-seq data. The green line indicates $G_g A_{g,z}$, the background fragment density derived from the observed data. The orange line indicates the model output without any contribution from S_s the sequence dependent bias component of the model. b) The single PCM that is the outcome from the model fitting.

Figure 3-8 indicates that there is a slight preference for the 5' end of an RNA fragment to start in the middle of a UA dinucleotide, and more generally for the fragments to start at nucleotides consisting of Us and As.

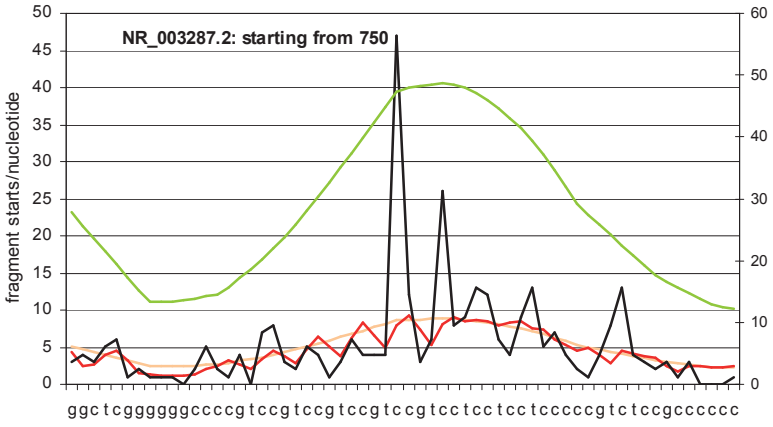


Figure 3-9 A region FRT-seq data from the 5' fragment end showing poor matching between observed data and model

Figure 3-9 shows a region of the 5' FRT-seq data where the match between model and observed data is poor, with a significant number of fragment starts at a location that is not predicted from the single PCM. There are a significant number of such locations in the genome where there are peaks in the sequence tag distribution that even a multi-PCM model was not able to match.

Figure 3-10 shows the PCM generated to model the sequence dependent fragment end distribution at the 3' end of the RNA fragments. This again shows that although the fragments are more evenly distributed through the genome than is seen with traditional RNA-seq protocols, there is still some nucleotide bias. While the bias at the 5' end was essentially restricted to the nucleotide on either side of the location of the end of the fragment, the bias at the 3' end appears to extend for four or five nucleotides into the sequenced fragment.

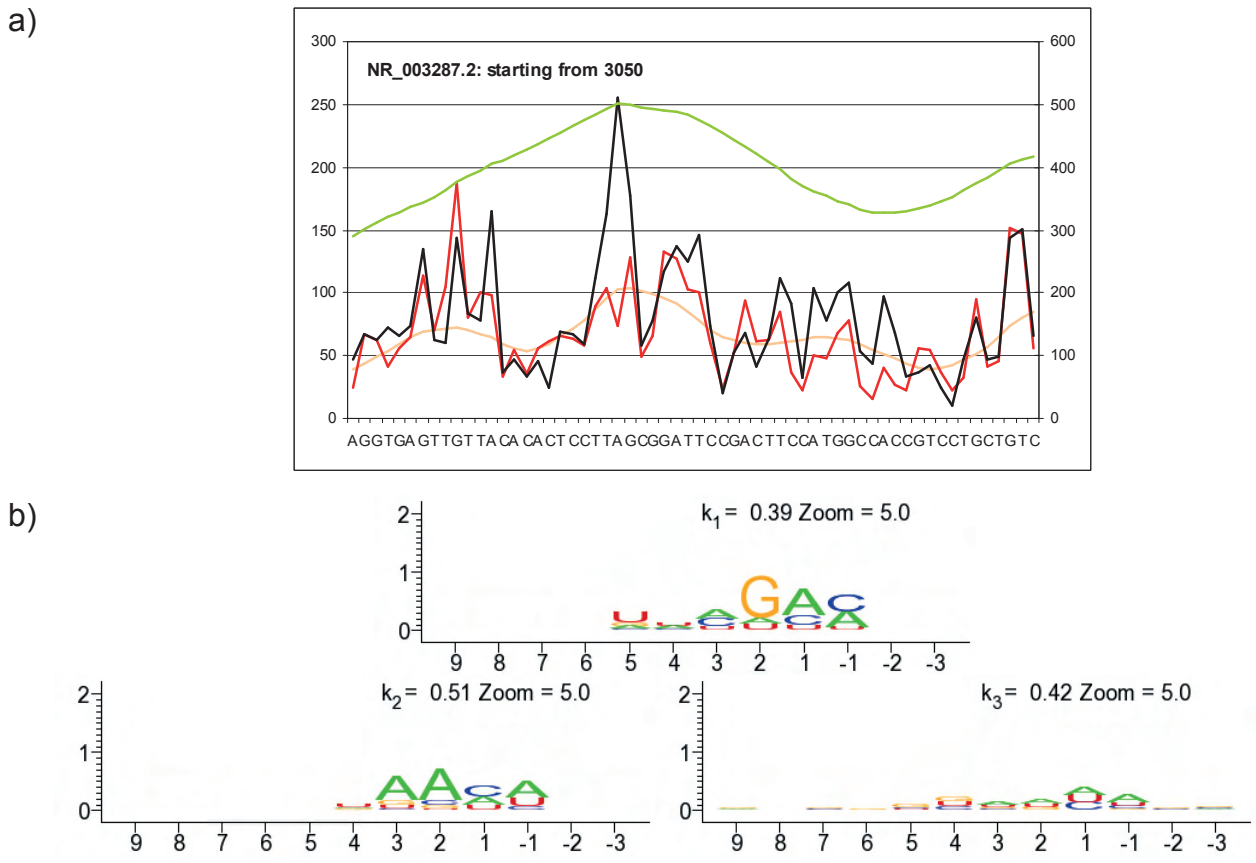


Figure 3-10 Model fitting of FRT-seq data 3' fragment end. a) Comparison between observed (black) and modelled (red) 5' fragment end distribution for a sample of FRT-seq data b) The three PCMs that forms the basis of the model-fitting. Positive numbers indicate the nucleotide within the sequenced fragment. Negative numbers indicate proximal nucleotides.

3.4 Discussion

3.4.1 Two distinct bias regions in RNA-seq data indicate two distinct molecular mechanisms †

Unlike the ChIP-seq data, the RNA-seq data examined show significant bias only on the nucleotides in the fragment itself and negligible bias in the nucleotides immediately preceding the start of the fragment. It has already been proposed [40] that this bias is due to the binding of the random hexamers to the RNA which constitutes an early stage in the conversion of RNA to DNA with reverse transcriptase.

However, the modelling in this paper suggests that the situation is more complex than was revealed in the analyses in previous papers, which assumed a single bias pattern. Modelling using multiple PCMs shows that there are two clearly distinct regions. The first, which requires multiple alternative PCMs to fit the observed data, covers the first six nucleotides. The second covers the region from nucleotide seven onwards, where a single nucleotide bias is observed which is virtually identical in all the data examined. This suggests that there are two distinct mechanisms that are responsible for the PCM patterns in these two regions. This may provide more information on the way that the random hexamer binding causes the observed bias.

3.4.2 Random hexamer related RNA-seq bias in nucleotides 1-6 †

The six nucleotide width of the first region is consistent with the hypothesis that the bias occurs as a result of the binding of the six-nucleotide-long hexamers. On previous occasions when a single bias was assumed, it was not possible to explain the results in terms of binding energies [40]. This new more complex insight into the binding should provide a better starting point for an examination of how DNA/RNA binding energies could give rise to the observed characteristic.

The asymmetry of the pattern in these six nucleotides is particularly striking, with a strong GC preference at the nucleotide at the 5’ end of the RNA. This would arise during the creation of the second strand of the DNA and may be an indication that binding is initiated at the 5’ end of the random primer. In addition, a preference for an initial GC binding may indicate that it is the three hydrogen bonds in this pairing, rather than the two-bond AT pairing, that makes it more likely that the binding will start with a CG pairing. The pattern for the following nucleotides shows a significant correlation between adjacent nucleotide positions, with a tendency for runs of Us, Cs or As. The pattern of runs of alternative

nucleotides was hidden in the previous analyses as a result of the assumption that there was only a single bias pattern present.

Appendix D also shows in more detail the virtually identical patterns for the 5' and reverse complement 3' end of the RNA fragments that have previously been observed. The bias at the 5' end will result from the binding of random DNA primers to the RNA as part of the process of creating the first strand of DNA, and the bias at the 3' end is a result of random primer binding to the DNA in order to create the second DNA strand. The more detailed model showing that the biases from both stages are very similar suggests that the random hexamer binding to DNA and RNA is governed by very similar physical processes.

3.4.3 Reverse-transcriptase related bias from nucleotide seven onwards †

The consistency of the pattern of nucleotides from nucleotide seven onwards suggests that it is caused by a different mechanism to that of the first six nucleotides. While these nucleotides may contribute to preferences in the binding of the random primer, they will be of greater significance when it comes to the binding of the reverse transcriptase and the processing of the enzyme along the RNA or DNA.

One possible explanation is that there is a greater probability of binding and transcription occurring if the first nucleotide after the random primer is an A and the fourth is a U/T. These data may consequently provide a useful additional insight into nucleotide preferences of the reverse transcriptase used in the RNA-seq protocol.

3.4.4 Implications for correcting bias in RNA-seq †

When correcting for any bias that is introduced in RNA-seq data, the existence of two separate mechanisms will influence the way in which the correction might be made. A preference for the random primer to bind at certain locations will affect the distribution of the start sites of fragments, but not necessarily the number of fragments that are ultimately sequenced in a specific region.

However, a bias in the likelihood that the reverse transcriptase will bind and transcribe a fragment may result in fragments in some regions being over or underestimated. This new analysis suggests that the details of any process for the removal of bias from RNA-seq data may depend on how these two effects might create inaccuracies in the characteristics being investigated.

Chapter 4

Protein binding site fingerprints in ChIP-seq data

This chapter builds on the evidence from Chapter 2 that ChIP-seq data contain information at a resolution of individual nucleotide positions and demonstrates how averaging techniques may allow information about the way proteins bind to DNA to be extracted from ChIP-seq data.

4.1 Introduction

Section 1.4 provides a general introduction to finding transcription factor binding sites using the ChIP-seq process. Section 1.4.10 provides an overview of the principles used by some of the peak finding algorithms which use the fragment distribution to locate protein binding sites.

A common factor of all peak-finding algorithms is that they do not use the tag counts at individual locations in the genome, which correspond to the number of fragments which start at that location. The typical approach is to divide the genome into segments that are 20 or more nucleotides wide and use the total tag count for each segment. One of the reasons for doing so is to reduce the number of data points down to a more manageable size, as working with a dataset containing an entry for every base-pair in the genome is computationally unwieldy.

Another reason for working with averaged data is the assumption that there is no significant information in the counts for each genomic location, with the local variation in counts from one nucleotide position to the next being due to random sampling effects and irrelevant artefacts that originate in the processing of the fragments.

This chapter describes a technique that demonstrates that information is available from ChIP-seq data at a per-nucleotide level. The technique extracts more detailed information about the nature of the binding between the protein and the DNA by making use of the way that the binding affects the likelihood that the DNA will fracture at specific locations during the fragmentation stage.

This is represented in Figure 4-1, which shows the motif associated with a protein binding site, and the pattern of fragment starts and ends in the region (see Section 1.4.8 for definition of fragment orientation). This chapter describes a method for analysing the data

from all of the sites where the motif is found and using the information to find out more about the way that proteins bind to the DNA at these sites.

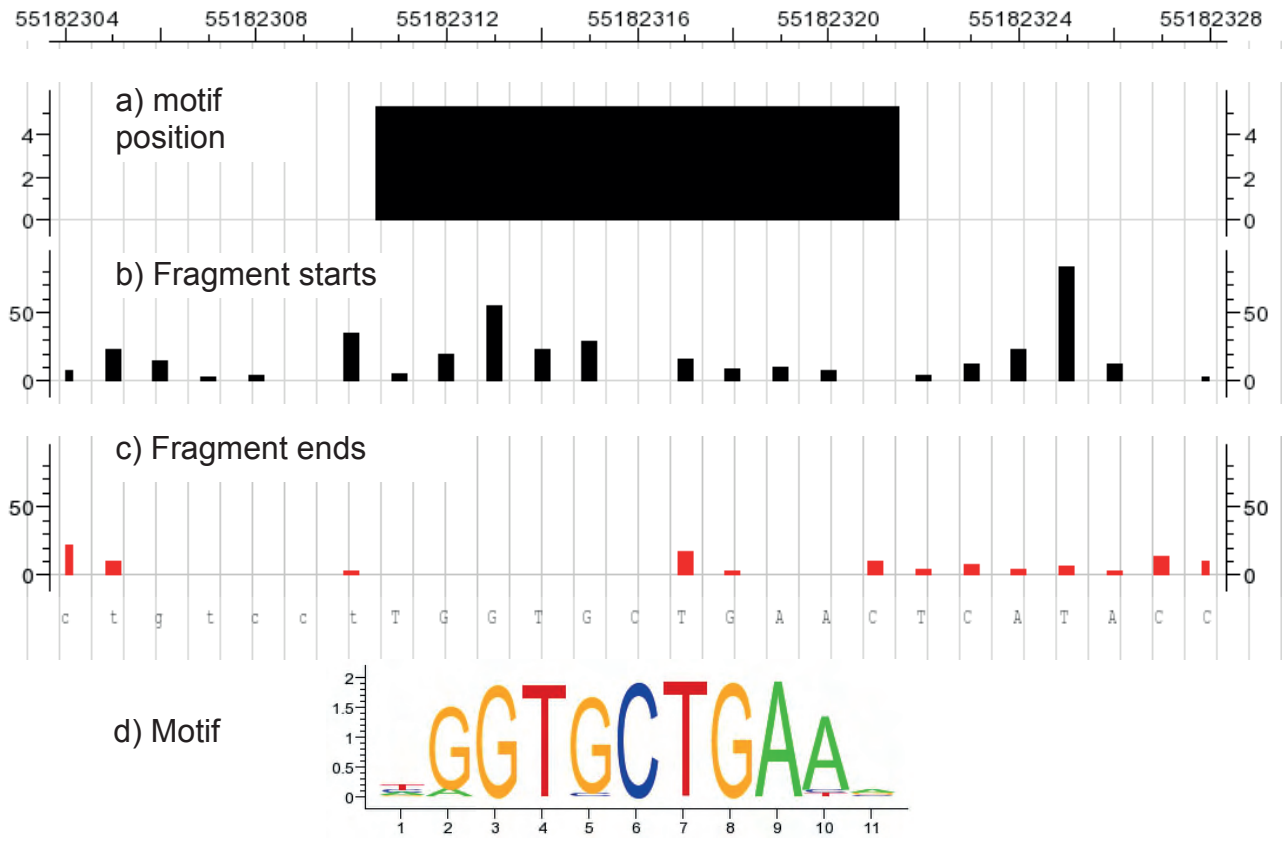


Figure 4-1 Relationship of fragment starts to motif. Region 55182304 to 55182328 chromosome 19 is shown which is in the region of an over-represented motif found in the SL523 data. a) Location of motif match as determined by cisGenome. b) numbers of fragment starts at each position c) numbers of fragment ends d) The motif, aligned to the genome.

It is sometimes the case that the binding motif for a transcription factor is not known at the start of a ChIP-seq experiment and one of the objectives of the experiment is to identify the likely binding motif. The techniques described then have a role in identifying which of the overrepresented motifs are associated with bound proteins, and whether one or more of the motifs are associated with the protein that was targeted at the immunoprecipitation stage.

4.2 Methods

The procedure adopted was to locate the peaks in the ChIP-seq data, examine the regions associated with the peaks for over-represented motifs, and then produce statistics that describe the relationship between the motifs and the distribution of fragment starts in the

region of the motifs. The cisGenome software used was a heavily customised version of cisGenome 1.0 (Appendix G-2).

4.2.1 Peak finding

Peak finding was performed using the peak finder that is integrated into the cisGenome software [47]. Part of its peak-finding algorithm involves comparing the enrichment of the immunoprecipitated fragments with that of the input DNA (Section 1.4.9).

One problem with the approach of using the ratio of the input and immunoprecipitated fragment density is that the fragment distribution in the input DNA can be very sparse. There may be only one or two fragments in the window size that is being used to calculate the ratio between the values from two datasets. This can result in significant variation in the ratio arising from numerically small differences in the number tags in adjacent windows.

For example, if there are three samples in one window and one sample in the next as a result of the random distribution of the sample sites, and these are used to measure the enrichment or fold change of the immunoprecipitated data then this can give the misleading impression that there is a three fold increase in the enrichment in the region associated with the second window compared to that of the first. This artefact was found to introduce significant skew in the predicted locations of some binding sites. An enhancement was introduced to the cisGenome software which allowed the input or background fragment density to use a larger window size than the signal data to reduce the effect of low input fragment density.

'Two sample' peak finding was performed which locates the regions where has been a significant increase in the fragment density compared to that of the input DNA. The default parameters (Windows size = 100, window step size = 25) were used. The window size used for the input DNA was 400. The value of p0, a cisGenome parameter indicating the relative density of the input DNA and the immunoprecipitated DNA, and the minimum read number that must be found in window before the peak is registered was determined using the exploration stage of the peak finding process provided by cisGenome.

The boundary refinement option in cisGenome was enabled so that the peak regions identified do not encompass the whole region of the peak but instead was narrowed down to the central region of the peak where the binding site is most likely to be located, using information derived from the distribution of the forward and reverse reads.

4.2.2 Identification of over-represented motifs

The Gibbs motif finder that is integrated into cisGenome was used to find the overrepresented motifs within the top N peaks, where N was chosen such that there were sufficient peaks to ensure that there was a sufficient variety of peaks within the sample whilst not resulting in an unduly long time taken to identify the motifs. The motifs are described using PCMs which give the likelihood of finding each specific nucleotide at each position. The width of the peak region previously identified by cisGenome was extended by 50 nucleotides to allow for the discovery of a range of over-represented motifs in the region of the binding site.

4.2.3 Identification of motif matches in the vicinity of peaks

The cisGenome motif mapping function was used to locate all of the instances within the region of the peaks where selected over-represented motifs matched the sequence. At each position the likelihood of a match between the motif and sequence is calculated and is compared with the same calculation using a third order Markov model of the genome (See section 1.4.11). A record is made of all locations where the likelihood ratio is 10 or more, together with the likelihood ratio at that location.

4.2.4 Calculation of fragment start fingerprints in the region of motif matches

The locations where there is a match between the motif and the DNA sequence in the region of the peaks were then used to identify if there are any specific patterns of fragment starts that are seen in the regions of these motifs. The general approach adopted is to find all the genomic locations where there is a degree of match between the sequence at that point and motif, and then find the average fragment distribution for all of these locations.

However, rather than find the overall average, the sites are grouped by the degree of match between the sequence and the motif, so an average is found for all the locations within a certain range of match to the motif at location x, as defined by the log likelihood L_x. of the sequence match at that position (see Section 1.4.11).

Let f_y be the number of fragment starts at the genomic position y, and $\mathbf{f}_x$ be the vector of fragment starts in the region of genomic position x of length P from x-a to x+b, i.e.

$$\begin{aligned} \mathbf{f}_x &= \left(f_{x-a}, \ldots, f_{x-1}, f_x, f_{x+1}, \ldots, f_{x+b}\right) \\ P &= a + b + 1 \end{aligned} \tag{4.1}$$

x is the position of the start of the motif in genomic coordinates of the strand to which the motif sequence matched.

A vector $\mathbf{F}_x(m,n)$ can then be created by summing all of the vectors for binding sites with $\log_{10}$(likelihoods) L_x in the range m to n. The vector $\mathbf{F}_x(m,n)$ is then converted to a normalised vector $\bar{\mathbf{F}}_x(m,n)$ where all of the values within the vector are scaled such that their mean value is one:

$$\begin{aligned}\mathbf{F}(m,n) &= (f_{-a},\ldots,f_{-1},f_0,f_1,\ldots,f_b) \\ &= \sum_{L_x>m,L_x<n} \mathbf{f}_x \\ \bar{\mathbf{F}}(m,n) &= \mathbf{F}(m,n)\times\frac{P}{\sum_{f_i\in\mathbf{F}(m,n)} f_i}\end{aligned} \tag{4.2}$$

The normalisation process ensures that the magnitude of the vector is independent of the number of locations that contribute to the vector and so allows the comparison of the vectors that have been produced with motifs with different ranges of log likelihoods. If the fragmentation probability is independent of the presence of the motif then the value of each coordinate in the normalised vector will tend to one.

A similar process can be used to create a vector $\mathbf{r_x}$ of the fragment ends, which are the fragment starts with respect to the opposite strand to the strand whose sequence matched the motif. The set of $\mathbf{r_x}$ can then be used to create the normalised vector $\bar{\mathbf{R}}_x(m,n)$ giving the characteristic distribution of fragment ends within the region of the motifs.

This approach results in locations where there are lots of fragments, such as those in the region of a very significant peak, making a significant contribution to the final vector compared to other locations where there are very few fragments.

The vectors created this way can create a picture of the fragment start distribution from all of locations where the motif matches the DNA sequence and where there is sufficient tag density data to be able to derive meaningful averages.

4.2.5 Normalisation of fragment distributions

The simplest option for creating the break distribution associated with a specific motif is to use the raw information about the location of the fragment starts that is generated by the ChIP-seq process. However, Chapter 2 showed that there is a genome wide sequence bias associated with the location of the fragment starts which tends to be specific to each set of data. Section 2.3.10 explored ways in which the bias can be compensated for in order to investigate other factors that contribute to the fragment distribution that is seen.

Such techniques can be used to compensate for this bias before investigating any bias associated with over-represented motifs. Two techniques were identified and both were used to remove the bias (Section 2.2.10). The first is to use the sequence bias values derived from the fragments that were not in the immediate vicinity of the peaks in the fragment distribution, and the second is to do model-fitting with this data, and then use the bias predicted by the model to adjust for the bias. The advantage of the second approach is that it may overcome problems with false biases that are generated for some sequences where there is a very low sample size such that the result generated from the original sequence data would be unduly affected by random variations in the number of counts

4.3 Results

4.3.1 Peak and motif finding from the NRSF immunoprecipitated SL522 dataset

The SL116 and SL522 datasets (Section 2.2.1) were selected to study whether there was a consistent pattern to the fragment starts and finishes in the region of potential binding motifs within this data. Both of these datasets consist of ChIP-seq fragments that were immunoprecipitated to select for bound NRSF protein. Two-sample peak finding was performed on the SL522 dataset using the SL523 dataset as the reference control (Figure 4-2). 11510 peaks were found, and the top 1000 were extended on each side by 50 nucleotides and cisGenome Gibbs motif finder used to identify the 10 most over-represented motifs (Figure 4-3).

A number of the motifs that were found, such as b), and j), are characteristic of the repetitive motifs that are frequently found in DNA. Motifs c) and i), and possibly a), d) and h) are more informative, and have more of the character of a protein binding motif.

Fragment start footprints for the motifs that were found were calculated using the fragment distributions for the top 2000 peaks. The first motif to be considered was the CCCC--CCC motif (Figure 4-3a). The fingerprints show the average fragment start density in the region of all of the motifs that occur within these peaks. The averages are grouped by the degree of match between the motif and the fragment so, for example, the results from all the locations that are a good match to the motif (i.e. log likelihood > 5) are grouped together as are all the results where the motif match is poor (i.e. log likelihood in the range two to three).

The x axis of the graphs is labelled with a consensus sequence that is derived from the DNA sequences associated with motif locations. The set of sequences used to create the consensus are those associated with the best match to the motif and where the group size is

greater than 20. A consensus nucleotide is defined as a nucleotide that occurs in more than half of the sequences. The match between consensus and motif is a confirmation of the presence of the motif at the centre of the sequence. The consensus can also give an indication of possible sequence conservation in adjacent nucleotides.

The SL522 results for the CCCC-CCC motif show considerable variation in fragment density across the region of the motif, and the variation is consistent between regions where the motif match is poor and where the motif match is good (Figure 4-4a).

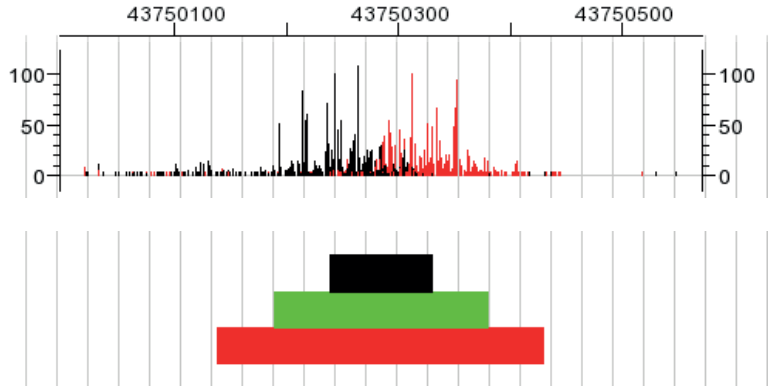


Figure 4-2 Example peak from the SL522 dataset. a) Fragment start (black) and finish (red) distribution for the nucleotides 43750011-43750557 in chromosome 22. b) Peak location as determined by cisGenome (black) and then extended by 50 (green) and 100 nucleotides (red).

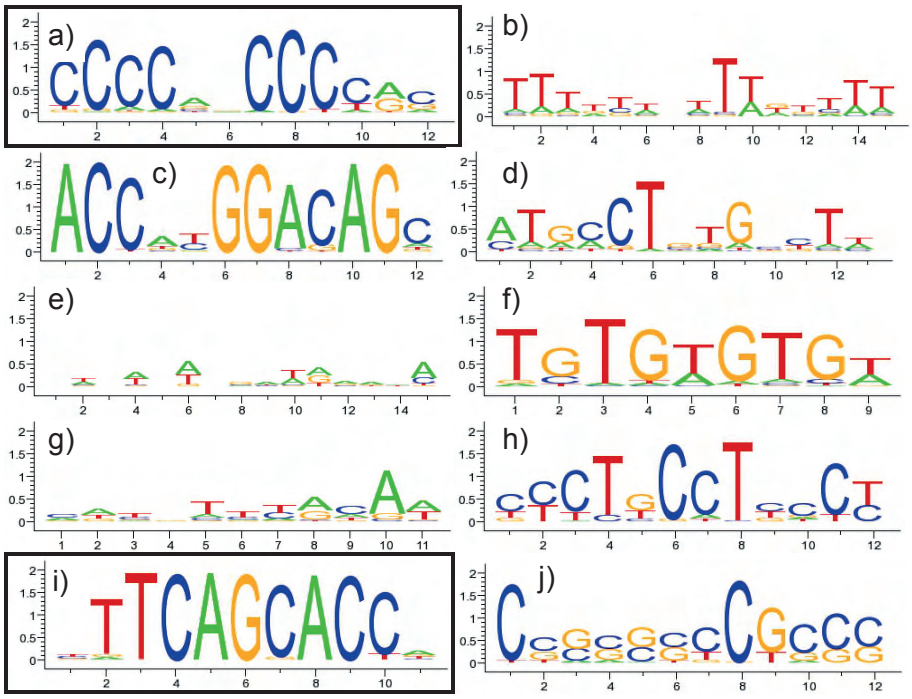


Figure 4-3 Ten Overrepresented motifs from the top 1000 peaks in the SL522 dataset. Motifs a) and i) are used in the subsequent analysis. This analysis showed that c) is a continuation of i) from position 8 and corresponds to the known binding motif for the NRSF protein.

4.3.2 Adjusting for sequence bias makes a significant difference to the binding fingerprint

It has previously been shown that there is a genome wide bias to the distribution of fragment start sites in ChIP-seq experiments (Chapter 2) and methods were designed to compensate for this bias in immunoprecipitated data using fragment data that did not come from the ChIP-seq peaks (Section 2.3.10). In order to assess the impact on the motif fingerprints of making such a correction, the CCCC-CCC fingerprints were generated using the fragment data after it had been corrected for sequence bias using the two different methods that had been proposed (Figure 4-4b and d). The results are very similar using the two techniques, and in both cases the significant variation in tag distribution seen when using the raw data largely disappears, leaving a distribution that is essentially flat, but with a slight depression in the region of the motif.

These results show, for one motif and one dataset, that compensating for sequence bias can have a significant effect on the fragment fingerprint of a motif.

4.3.3 Adjusting for sequence bias improves the alignment between fingerprints from different datasets

Given the effect on the fingerprint generated from the SL522 data of bias adjustment, the equivalent fingerprints were generated from the SL116 dataset to determine to what extent the fingerprints might match and also to determine the effect of sequence bias compensation (Figure 4-5). The sequence bias characteristics for the SL116 dataset are very different from the SL522 sequence bias, so the effect of bias compensation would be expected to be different.

The peaks for the SL116 data that indicate the regions of NRSF binding match the SL522 data, so the fingerprints for the SL116 fingerprints were generated using the same set of motif positions as were used for the SL116 data.

The fingerprints generated using the raw data do not have the same distinctive feature as was the case for the SL522 data but are comparatively featureless (Figure 4-5a). However, after compensation for the sequence bias, the SL522 fingerprint is broadly similar to the SL116 fingerprint, with a shallow dip in the region of the motif, suggesting that it is appropriate to use sequence bias compensation when analysing motif fingerprints, in that it does appear to draw out a common underlying characteristic, even from datasets where there was significantly different sequence bias and fingerprint generated from the uncompensated data.

4.3.4 The NRSF motif fingerprint adds further support to the principle of correcting for bias

The fingerprints for the raw fragment distribution data for the over-represented motif 4-3i) were created for the two datasets (Figure4-6) and also with the fragment data after compensation for the sequence bias using the sequence bias of the fragments other than in the peaks (Figure 4-7).

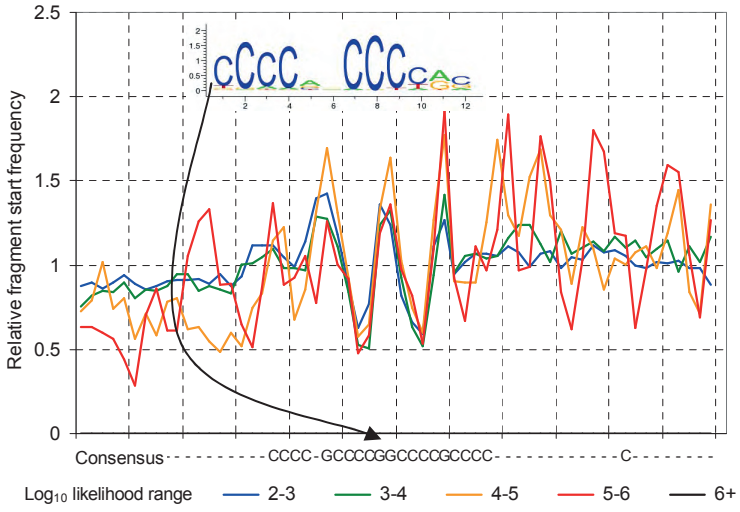
The fingerprints generated using the raw data are very different from each other, with the SL522 data showing considerable variation in tag density across the region corresponding to the motif. This mirrors the results seen previously for the CCCC--CCC motif. After compensation for sequence bias, the fingerprints from the two datasets are much closer to each other, adding further support for the validity of using sequence bias correction when looking at motif fingerprints.

The consensus sequence derived from the DNA sequences and used to label the x axes in Figure4-6 extends considerably to the right of the motif sequence with a characteristic GGACAG sequence. The combination of this together with the ACC at the end of motif 4-3i) matches the overrepresented motif 4-3c). The combination of these two motifs corresponds to the known binding motif for the NRSF protein [49], which was the target protein for these ChIP-seq experiments.

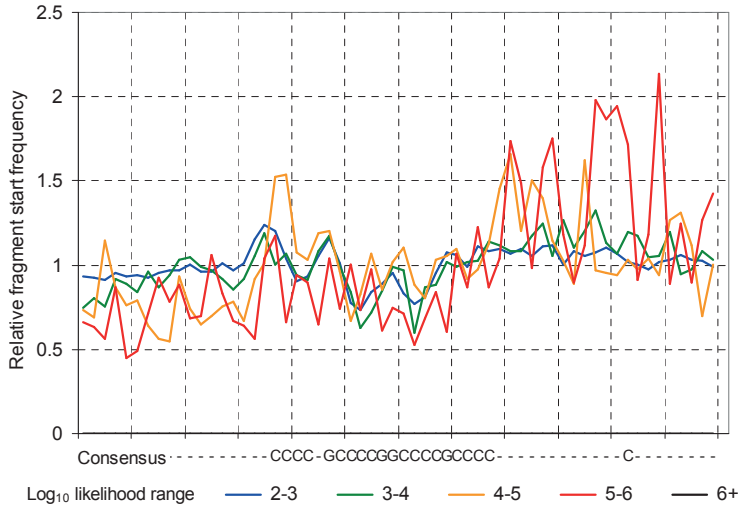
The fragment start distributions both show a very distinctive peak in the region immediately before the start of the region of DNA that matches the PCM.

In contrast to the results for the motif 4-3c) these results show distinctive slopes across the range of nucleotides that encompass the motif matching sequence. The fragment start fingerprints show a rising slope and the fragment end fingerprints show a falling slope.

a) No sequence bias compensation: SL522

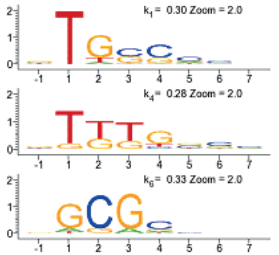


b) Sequence bias compensation using data from fragments



c)

log likelihood	Motifs
6+	39
5-6	234
4-5	1150
3-4	3861
2-3	37197



d) Compensation using bias predicted from model

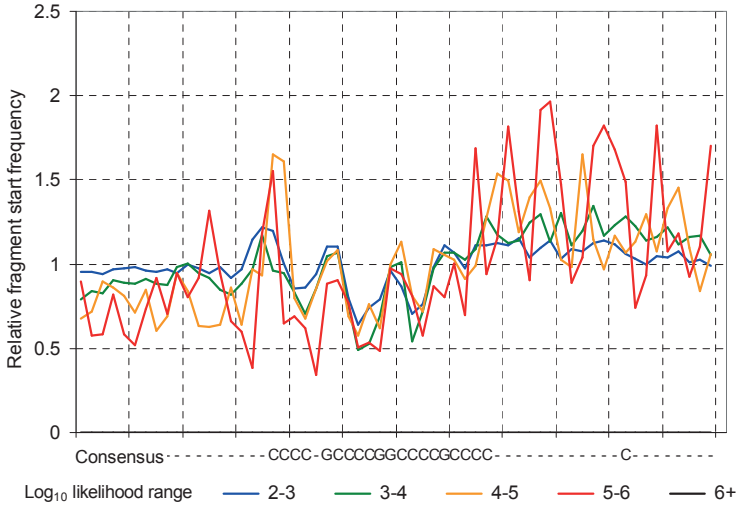
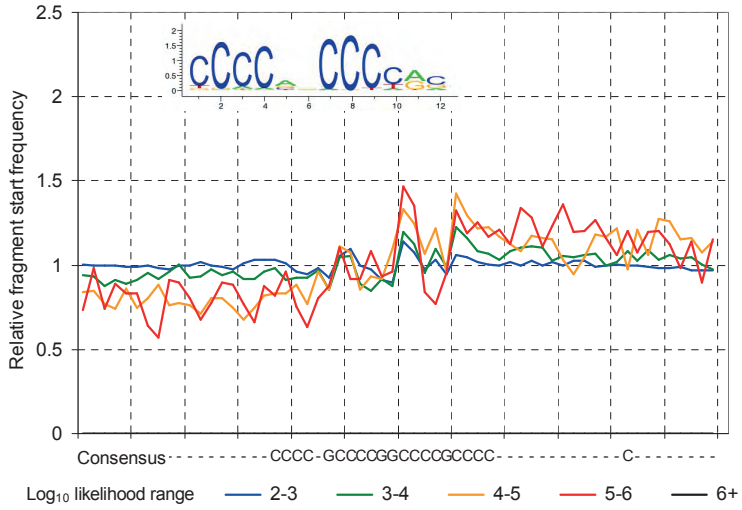


Figure 4-4 Fingerprint for CCCCXXCCC motif in SL522 dataset largely disappears after compensation for sequence bias.

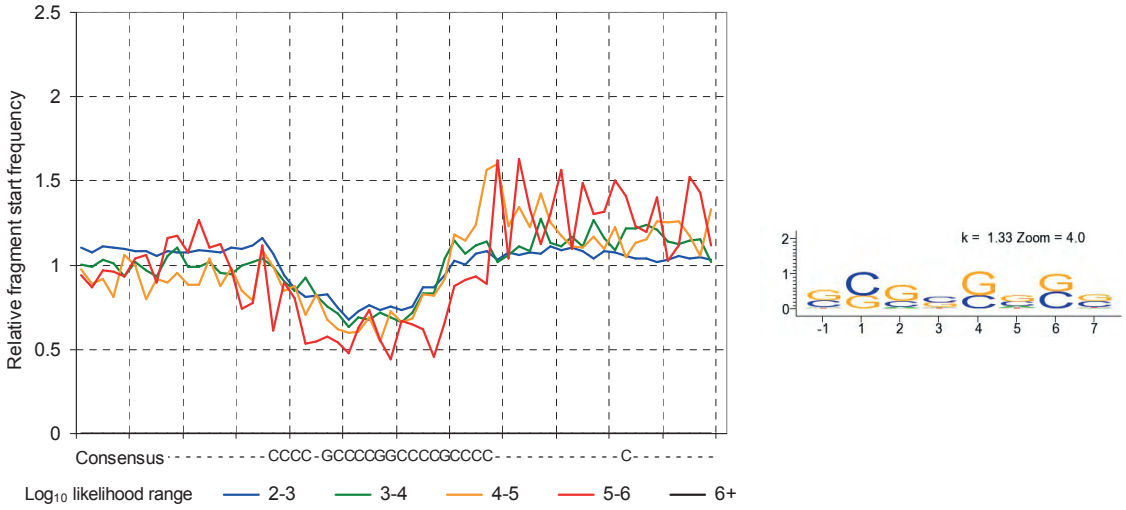
a) Fragment start distribution fingerprint calculated with raw fragment data. b) Distribution after compensation for sequence bias calculated using sequence bias of fragments other than in the peaks d) Distribution after compensation using bias predicted from model. c) Number of motifs in each log-likelihood range and sequence bias PCMs for this dataset. In a),b) and d) the regions are grouped based on the log likelihood match between the sequence an the motif, and average fingerprints calculated for each group.

The results obtained from the raw data would suggest that the DNA is more likely to fragment at some positions within the region of the motif compared to others. However, compensation for the sequence bias of fragment removes this apparent relationship. After compensation there is instead just a slight dip in the fragment start likelihood in the region of the motif

a) No sequence bias compensation: SL116



b) Sequence bias compensation using data from fragments



c) Compensation using bias predicted from model

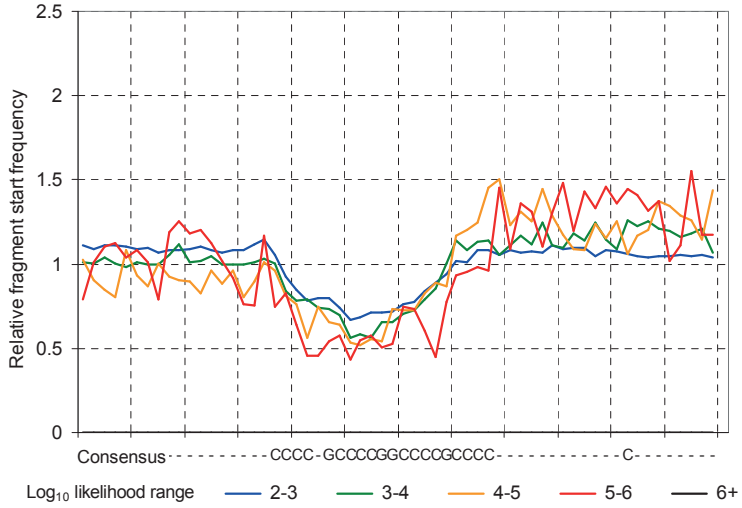


Figure 4-5 SL116 Fingerprint for CCCC−−CCC motif is similar to SL522 after compensation for bias.

a) Fragment start distribution calculated with raw fragment data b) Distribution after compensation for sequence bias using sequence bias of fragments other than in the peaks d) Distribution after compensation using bias predicted from model. While the fingerprint created using raw data is relatively featureless compared to the SL522 fingerprint, after compensation for sequence bias the fingerprints shows a slight tendency for there to be fewer fragment starts in the region of the motif. Given the tendency for fragmentation to occur in GC-rich regions and the C richness of the motif, an increase in the raw fragment density would be expected. This is not seen, so the compensated results show that fewer fragments are seen than would have been expected. This dip is similar to that seen in Figure 4-4.

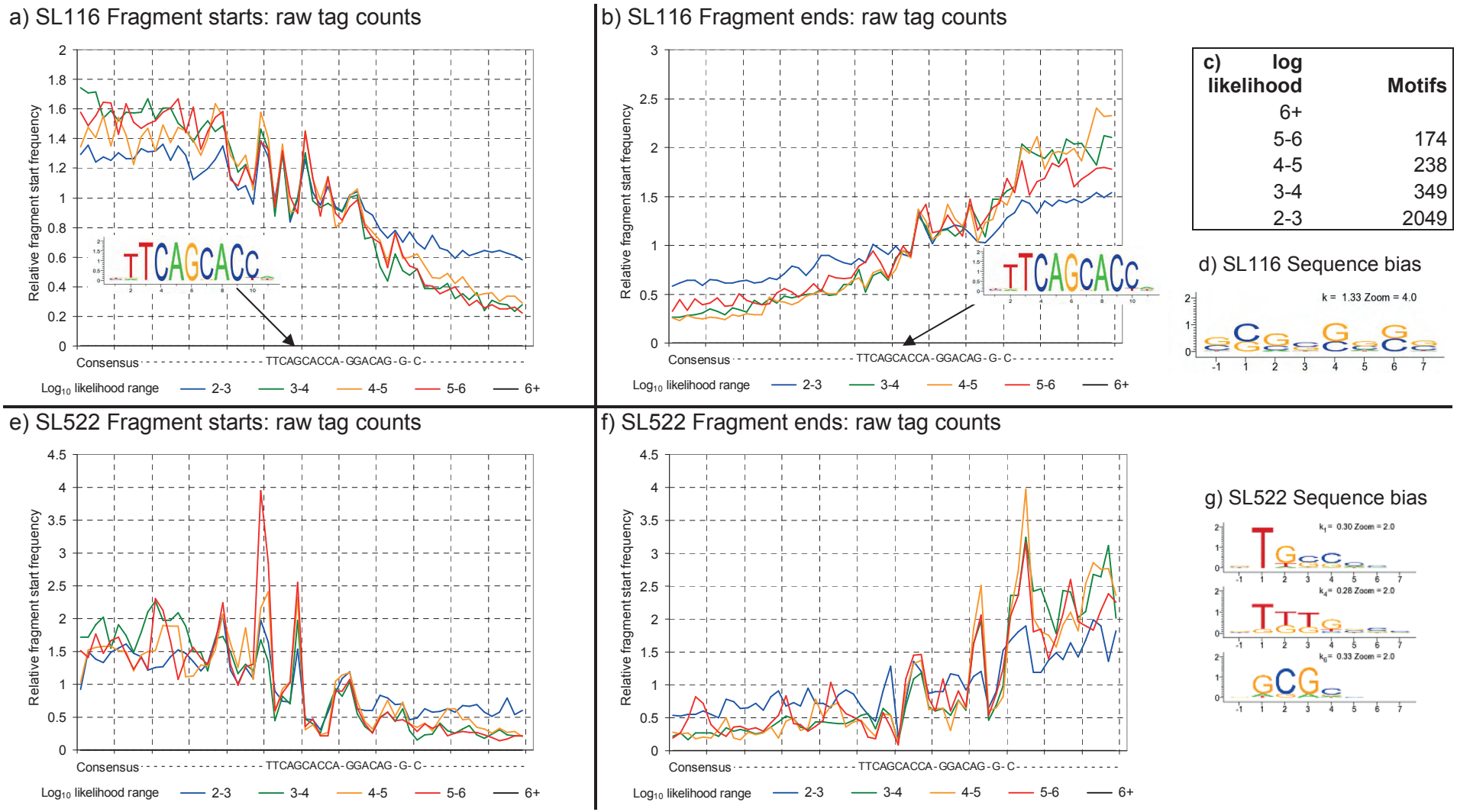


Figure 4-6 Fingerprints associated with NRSF binding motif and generated from raw counts from two datasets show very different patterns.

a&b) Fingerprints calculated using SL116 data, fragment starts and ends. c) Number of motifs in each log-likelihood range d) SL116 Sequence bias PCM. e&f) Fingerprints calculated using the same motif and peak definition for the SL522 data. g) PCMs representing SL522 sequence bias.

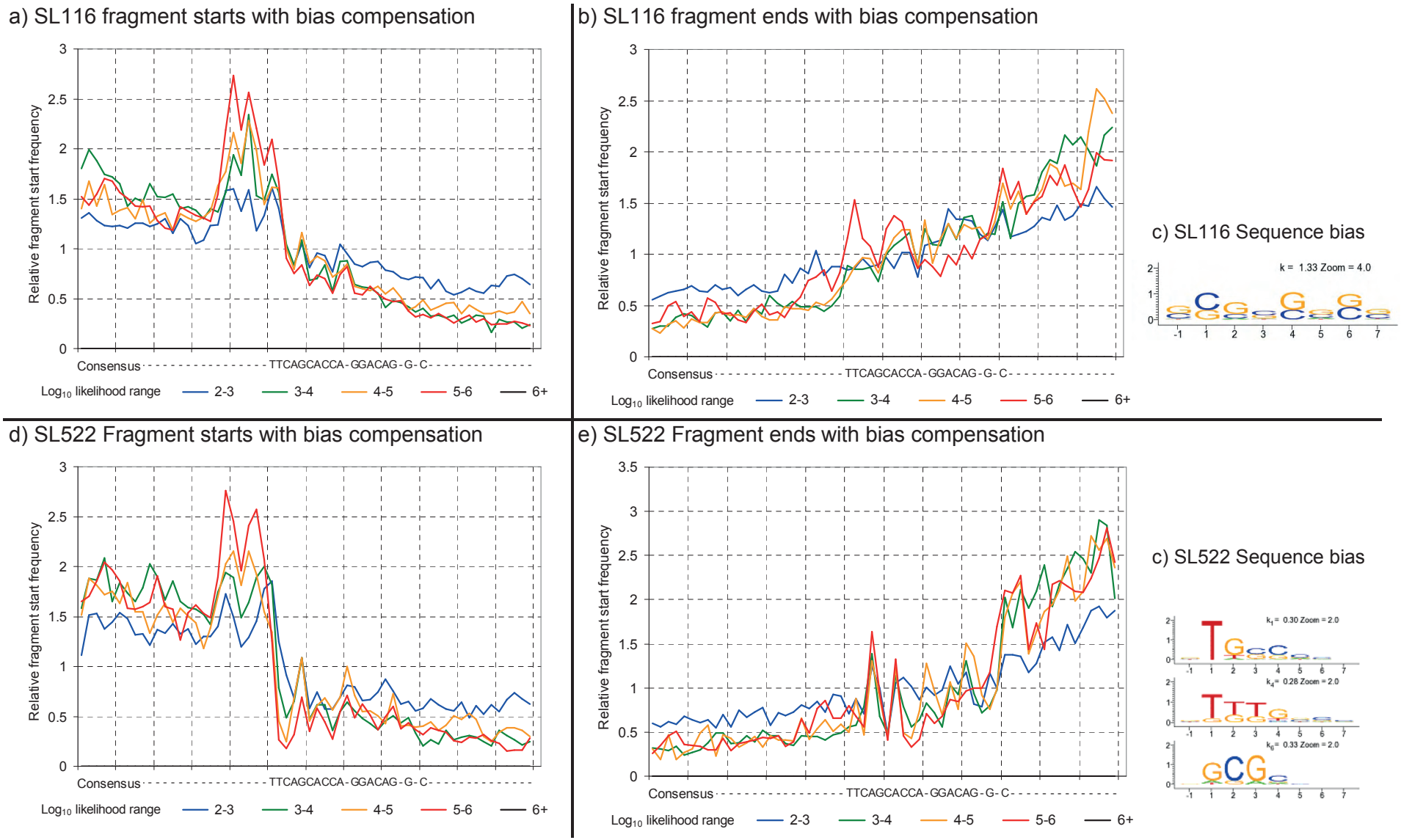
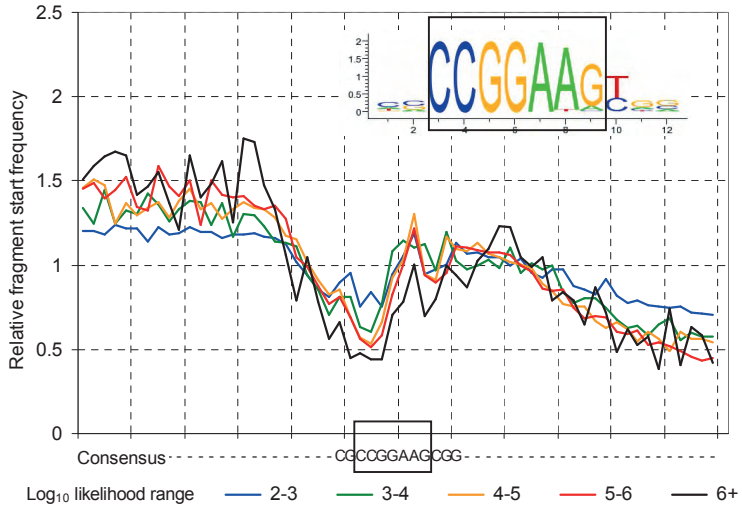
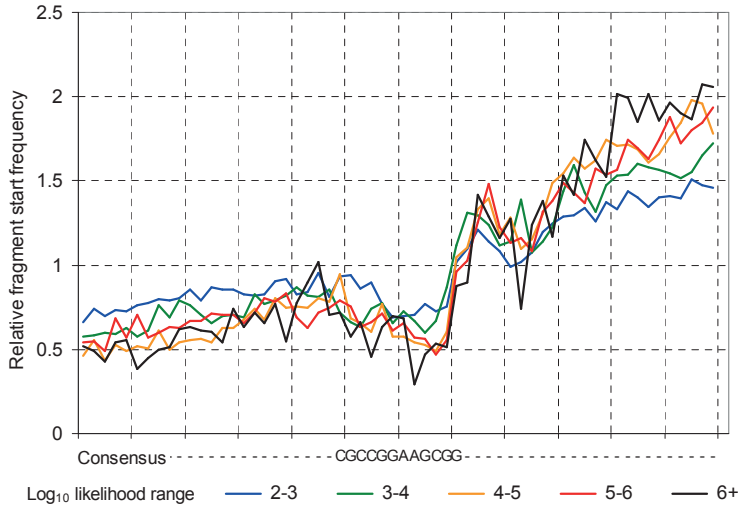


Figure 4-7 NRSF fingerprints from different datasets are similar after sequence bias compensation. a&b) Fingerpints calculated using SL116 data. c) Number of motifs in each log-likelihood range d) SL116 Sequence bias. e&f) SL522 fingerprints calculated using the same NRSF motif and peak positions as used for the SL116 data. g) PCMs representing SL522 sequence bias.

a) SL610: Fragment starts

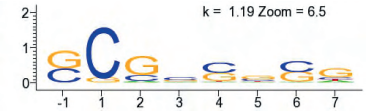


b) SL610: Fragment ends

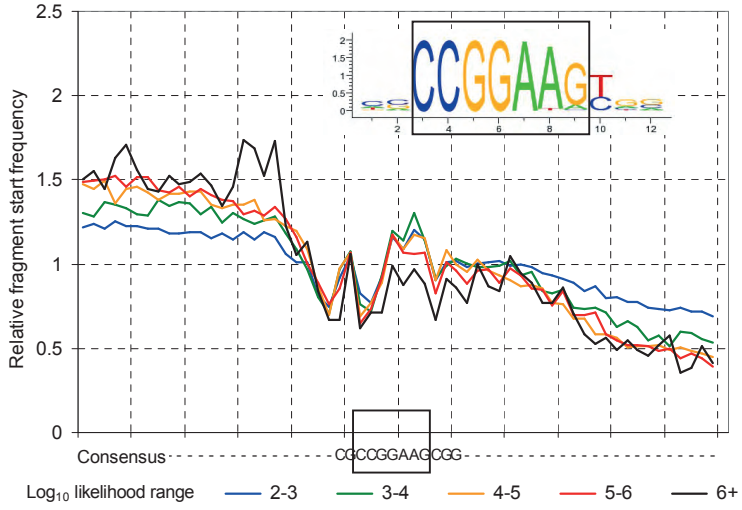


c) log likelihood	Motifs
6+	68
5-6	280
4-5	358
3-4	258
2-3	1267

d) SL610 Sequence bias



e) SL223: Fragment starts



f) SL223: Fragment ends

g) SL223 Sequence bias

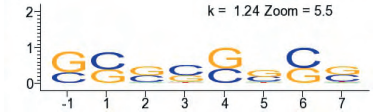


Figure 4-8 GABP fingerprints from different datasets show some similarity, but also significant differences. a/b) Fingerprint for the GABP motif from the SL610 dataset, corrected for sequence bias c) The number of motifs in each range of log likelihoods d) The SL610 sequence bias e/f) Fingerprint from the same locations in the SL223 dataset, also corrected for sequence bias g) The SL223 sequence bias, which is similar to SL610.

4.3.5 Poorer fingerprint match seen in GABP motifs from different ChIP-seq experiments

The SL610 and SL223 datasets come from fragments immunoprecipitated for the GABP transcription factor from HeLa and K562 cell lines respectively. A GABP binding motif was derived from the SL610 data and used to generate fingerprints for the two datasets (Figure 4-8). In both cases characteristic slopes across the 60 nucleotide region being examined can be seen which match the slopes seen in Figure4-6.

In addition, both sets of data show dips in the region of the motif itself, indicating that slightly fewer fragment starts are seen in this region. There are however distinct differences in the detailed characteristics that are seen.

4.4 Discussion

While some initial ChIP-seq fingerprints were obtained at the very start of the research, the versions of the results presented in this document only became available at the very end of the research period because of their dependency on the investigation into sequence bias that was covered in earlier chapters. Consequently it has only been possible to investigate a small amount of data using the methods that have been developed and any discussion presented is therefore very tentative. This is particularly the case in the light of the unexpectedly large variation in the sequence bias characteristics that have been seen in ChIP-seq data, making it unwise to draw too general a conclusion from a small set of these sorts of results (see for example Section 2.4.4).

4.4.1 ChIP-seq data contains information at a single nucleotide resolution

These results are perhaps the best support for the original motivation for this research, namely that the ChIP-seq data contains information at the resolution of individual nucleotides. While this was arguably true for the sequence bias results, the motif fingerprints appear to provide more specific support for this thesis. Examples of single nucleotide resolution include the rapid increase in the probability of a fragment end immediately after the end of the GGTGCTGAA motif in Figure 4-7b) and e), and the spikes in the fragment start distribution density in Figure 4-8.

4.4.2 Sloping fingerprints indicate motifs associated with the target protein

The fingerprints for motifs that are associated with the target proteins all show ‘shoulders’ or characteristic falling slopes for the distribution of fragment starts, and rising

slopes for the fragment ends (Figures 4-6,4-7 and 4-8). The explanation for this shape comes from considering the overall fragment distributions in the region of the location of the target protein. The shape arises because all the fragments overlap sites where the protein was bound in order to have been selected by immunoprecipitation. If there is only one such site in a region this gives rise to a very characteristic twin peak distribution for the fragment starts and ends (Figure 2-17). The fingerprint is an average of a small section from the middle of this distribution and will reflect the typical rising and falling distributions in this region (Figure 4-9).

There are two cases to consider for motifs that are not directly associated with the target protein. The first is where the motif location is unrelated to the position of the target protein, and so will occur at a variety of different positions in the twin peak distribution, and the average of the different slopes from the different positions will tend towards being flat. Consequently, the flatness of the fingerprint associated with the CCCC-CCC motif would suggest that its position is not significantly related to the location of the target protein.

The second case to consider is motifs that are associated with the location of the target protein, such as motifs that are associated with the binding of cofactors that tend to be immediately adjacent to the target protein. The fragment distribution fingerprints for such motifs will show similar shoulders to those of motifs directly associated with the target protein, because they are in the same position relative to the typical fragment distribution as the target protein. A more detailed analysis and comparison of the distributions may be able to distinguish between the different categories of bound protein.

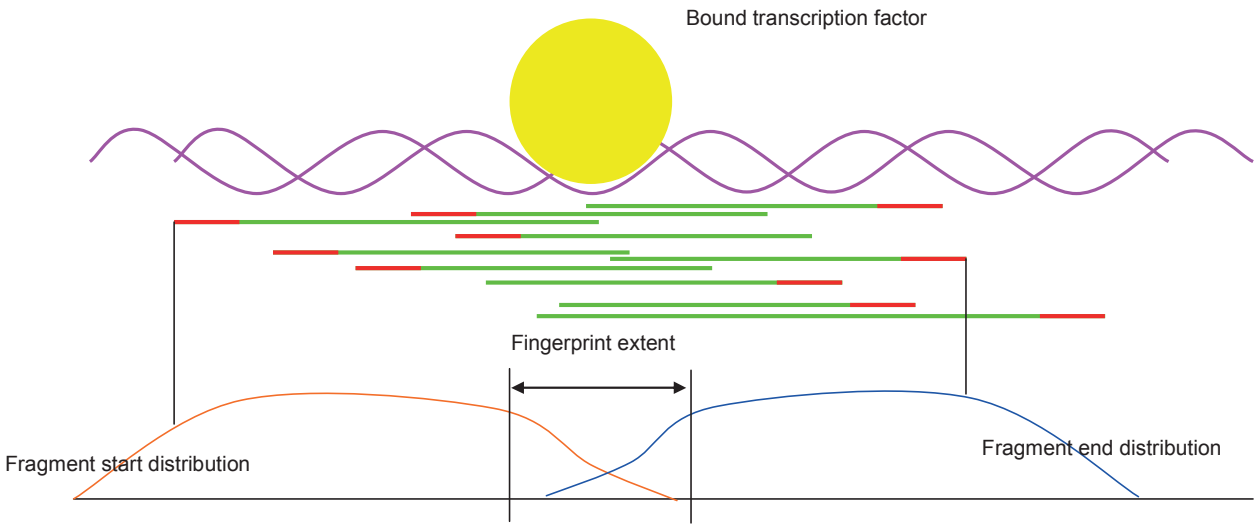


Figure 4-9 Origin of slopes in fingerprints for target proteins. Fragments (green) that are selected by immunoprecipitation will largely overlap the location where the target protein (yellow) was bound to the DNA. The starts of the fragments will tend to be located before the protein position, and the ends will lie afterward. The overall profiles for the fragment start and ends are shown in orange and blue. The fingerprints are an average of the region of these profiles that lie close to the binding site and will show a small section of the overall profile.

In the examples analysed the slope becomes less pronounced with poorer matching between the sequence and the PCM. In each graph the blue line corresponds to the set of regions with a log likelihood match of 2-3, the poorest range of matches considered, and its slope is always the least pronounced (Figures4-6 to 4-8). It is nevertheless the case that the blue line shows a significant slope, suggesting that binding is still taking place even with a relatively weak match between the motif and sequence.

4.4.3 Fingerprints may provide more detail as regards protein binding

When considering the sequence bias that is seen in ChIP-seq data (Chapter 2), the simplest explanations for the observed bias relied on assuming that the fragment start and end positions determined during sequencing gave an accurate indication of where the DNA originally fragmented during sonication. Variations from a flat distribution of fragment start locations would therefore indicate that something has affected the probability of the DNA fragmenting at these locations during sonication. One significant factor is likely to be the presence of a protein bound to the DNA (and fixed in position using formaldehyde), so the distinctive shapes of the fingerprints in the region of a motif could result from the effect of the presence of the protein on DNA fragmentation.

In some cases there are one or more dips in the fingerprint profile within the region defined by the motif, suggesting that the presence of the protein makes fragmentation at these specific locations less likely (e.g. in Figure 4-8).

In some cases (e.g. Figure 4-7a and d) the lines for the different likelihood ranges do not coincide, but become more extreme with each improvement in the match between the sequence and the PCM. If the degree to which the fingerprint becomes more extreme is roughly in line with the change in the slope then it suggests that as the probability of binding increases, the degree of effect that the presence of the protein has on fragmentation also increases.

However in the case of Figure 4-7 a and d, there is very little change in the slope between a log-likelihood range of 2-3 and the range 4-5, indicating no significant change in the degree to which the protein is bound to the DNA over this range of sequence matches. There are however significant differences in the sizes of the peak, suggesting that the effect that the protein has on DNA fragmentation is a function of the match between the DNA sequence and the canonical sequence.

4.4.4 Peaks in binding footprints may provide information on chromatin remodelling by NRSF

The binding motifs for most eukaryotic DNA binding proteins are asymmetric, indicating an asymmetry in the binding between the protein and the DNA. A clear asymmetry can also be seen in the fingerprints associated with the motifs examined (Figures 4-7 and 4-8). For example, the peak in the fingerprint seen immediately to the left of the TTCAGCACC NRSF motif is not seen on the right, or on either side of the GABP motif.

The TTCAGCACC motif identified during the process of model fitting is part of a longer canonical motif, the other part of which was also found when searching for over-represented motifs (Figure 4-3c). This can also be seen in the conserved sequence to the right of the primary motif in Figure 4-7. The peak in fragment start distribution is not associated with the other component of the NRSF motif, so is not associated with the core NRSF binding site but lies slightly outside this region.

A possible explanation for the distinctive fingerprint relates to the mode of action of NRSF as a chromatin remodelling complex which regulates gene expression by changing the local DNA folding. The increased level of DNA cleavage that is seen immediately adjacent to the binding motif could be as a result of the NRSF changing the DNA confirmation such that it is more vulnerable to shearing during sonication (Section 1.4.3). The use of ChIP-seq data

to investigate the effect of protein binding on the local conformation of the DNA would represent a new application of ChIP-seq data, although closely related to existing applications such as its more general use in investigating chromatin structure (Section 1.4.13).

4.4.5 Common features in two GABP binding fingerprints may indicate aspects of bond between DNA and GABP

The fingerprints for fragments immunoprecipitated for bound GABP from the SL610 and SL223 datasets show some similarities, but also significant differences (Figure 4-8).

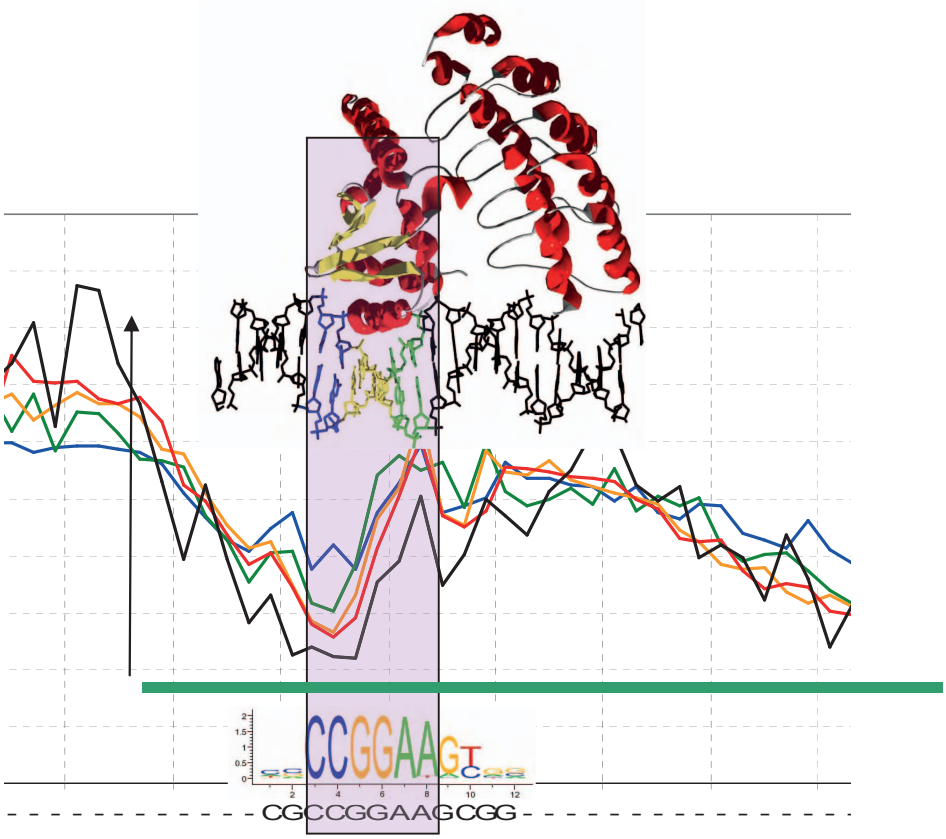


Figure 4-10 Fragment start fingerprint for GABP motif and SL610 dataset . The generating motif for the fingerprint is shown above the sequence consensus at the bottom of the graph. The structure of GABPα/β bound to DNA has been aligned so that the DNA sequence aligns with the motif, with equivalent nucleotides colour matched. A representative DNA fragment is shown as a horizontal green line with the fragment start aligned to the most probable fragment start position (arrowed) [8]. Fingerprints were calculated for different ranges of matches between the motif and the sequence log-likelihood values covered: blue (2-3), green (3-4), orange (4-5), red (5-6) and black (>6).

Both fingerprints for fragment starts show a dip in the fingerprint in the region around the start of the motif. The region was aligned to the known structure of GABPαβ bound to DNA (Figure 4-10)[8]. The dip is seen to extend for four or five nucleotides to the left of the

binding region as shown, suggesting that some aspect of the binding of GABPα to DNA reduces the likelihood of the DNA fracturing in this region. This could, for example, indicate the frequent presence of a co-factor.

The likelihood of the DNA fracturing increases through the region of the binding region between GABPα and the DNA, reaching a maximum at the two A nucleotides. This feature is seen in both datasets (Figure 4-8), suggesting that a significant aspect of the bond between GABPα and the DNA is responsible for this.

4.4.6 There are significant differences between the two GABP binding fingerprints.

As well as common features, the two GABP datasets show distinct differences. The SL223 dataset shows well defined spikes at specific locations for both the fragment start and finish distributions which are not seen in the SL610 datasets (Figure 4-8). The spikes appear to be significant in that virtually the same shape is seen for the five different sets of regions, where each associated with a different range of log likelihood matches between sequence and PCM so represent independent sampled from the ChIP-seq data.

A potential source of variation between experiments is the differences in experimental conditions that result in differences in sequence bias that are seen between experiments. However the sequence bias for the two sets of data show the very similar GC-rich characteristic making it less likely that this is responsible for the differences that are seen (Figure 4-8d and g).

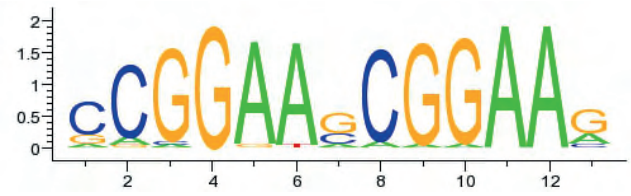


Figure 4-11 Over-represented dual GABP binding motif found in binding peaks of SL223 dataset. This suggests that GABP may bind in the known hetero-tetramer formation and this may need to be taken into account when interpreting the fingerprints.

One other aspects of these results that may need to be considered in order to better understand them is that one of the over-represented motifs found in the SL223 dataset consists of two closely spaced GABP binding motifs (Figure 4-11), and GABP is known to form a stable heterotetramer consisting of two GABPα and two GABPβ subunits [21]. It is possible therefore that some aspects of the fingerprints that have been found are associated with

heterotetramer binding, and there may be a degree of variation of tetramer formation between experiments, giving rise to the differences that are seen.

Chapter 5

Using modelling to study SeqA binding in E. coli

This chapter shows how the modelling techniques used in Chapter 2 and Chapter 3 can also be used to obtain more detailed information from ChIP-chip data about how proteins bind to prokaryotic DNA.

5.1 Introduction

5.1.1 The role of SeqA in prokaryotic cell replication

Eukaryotes and prokaryotes both use extended DNA molecules to encode the information required to assemble proteins, and both use similar processes of transcribing DNA to RNA using RNA polymerase followed by the assembly of the protein sequence using ribosomes. However they differ markedly in the way in which these processes are controlled. This is not unexpected given the challenge faced by the eukaryotic cell as a result of its need to use the information in the genome to change the cell programming in far more complex ways than is the case for prokaryotes. An example of this additional complexity is the challenge of switching the genetic program as a function of the cell type in multicellular organisms.

However both eukaryotes and prokaryotes use some common mechanisms to implement the required programming, including the control of gene expression through the binding of proteins to DNA. In the case of eukaryotes, this is frequently performed through the binding of proteins to regulatory regions, which are extended regions immediately upstream of the transcription start site, which then control gene expression through complex and often still poorly understood interactions.

In prokaryotes the control mechanism, although still complex is considerably simpler. One example of the modulation of gene expression through protein binding is the binding of SeqA to the genome as part of the process of controlling cell replication [97]. Its role in replication was originally discovered in *E. coli* [67], although the conservation of the distribution of GATC sites suggests that its role is widely conserved across prokaryotes [86].

During cell division it is important to ensure that replication is not reinitiated on the newly replicated DNA. It was discovered that GATC sites are hemimethylated after the DNA has been replicated [18] and that SeqA binds to these sites co-operatively [38]. At these sites, it then prevents the reinitiation of replication, particularly as a result of being bound to the

origin of chromosomal replication (oriC) site which contains multiple GATC sequences. SeqA is a 21 kDa protein which forms a homotetramer which can bind to two hemi-methylated GATC sequences separated by up to 31 bases [39]. In addition, it has been suggested that SeqA binding to GATC sites in the *dnaA* promoter region prevents the expression of the *dnaA* gene which acts as an additional barrier to premature reinitiation of replication [67].

There are indications that SeqA may play a wider role in gene regulation in that less SeqA is found in highly transcribed regions and SeqA remains bound to the genome for some time after such binding would have an obvious role in the control of replication [86]. This is consistent with other evidence that it may play a wider role in transcription regulation [90]. In order to understand more fully how it might fulfil such a role ChIP-chip was used to quantify the degree of SeqA binding to the *E coli* genome at various points in the cell cycle [86].

This chapter describes the results of a re-examination of these ChIP-chip data in order to extract more detailed information about the *in vivo* binding of SeqA to the *E. coli* genome. These results confirm details of the binding cooperativity which was previously identified from *in vivo* experiments [15]. They also provide new information showing that the probability of binding depends on the two nucleotides on either side of the core GATC binding motif.

5.1.2 Applying modelling techniques to ChIP-chip data

The ChIP-chip data that had been produced to investigate the binding of SeqA to the *E. coli* genome provided an opportunity to explore whether the modelling techniques developed to model DNA and RNA fragmentation could also be used to draw more information out of ChIP-chip data. The underlying principle is, as was the case in previous chapters, to create a model that attempts to reproduce aspects of what is believed to be happening at a molecular level. Model fitting is then used to determine the parameters of the model and from these parameters understand more about what is happening at a molecular level.

5.2 Methods

5.2.1 Preparation and choice of ChIP-chip data

E coli K-12 cells with a with a temperature sensitive *dnaC2* mutant were used as the source of DNA for the ChIP-chip fragments. The mutant allows the cells to be arrested immediately prior to replication so that the cell division cycle can be synchronised. ChIP-chip analysis was performed on unsynchronised cells, cells in the blocked state and cells six

minutes after they had been released from the blocked state. The preparation is described in more detail in the associated published paper [86].

The analysis of SeqA binding was carried out using the ChIP-chip data from unsynchronised cells because these data provided evidence about the degree of SeqA binding throughout the genome, whereas SeqA binding in synchronised cells was largely restricted to specific regions. In providing genome wide data there was data over a greater number and variety of SeqA binding sites, which gives the model fitting process more information to work with.

5.2.2 Principles of modelling SeqA binding

The immunoprecipitation stage of the ChIP-chip procedure selects DNA fragments to which the target protein is bound at least once. The range of lengths of these fragments is determined by the previous sonication stage. After the protein is removed from the fragments, the fragments are amplified, tagged with a fluorescent tag and allowed to bind to a microarray with complementary probes spaced regularly through the genome (see Section 1.4.6.) The fluorescence at each point can be used to determine the quantity of the target protein that was bound in the region of each probe.

The challenge for any mathematical model of this process is that the final data are only a very indirect measure of the degree of protein binding. Most of the stages in the ChIP-chip process will have some effect on the level of binding at a particular probe site as seen thorugh the measurement of the binding using fluorescence. The effect of these stages muts therefore be incoporated into the model.

5.2.3 Modelling of the effect of adjacent dinucleotides on SeqA binding

The simplest model for SeqA binding is that $P(B_x)$, the probability that SeqA will bind at position x can be represented by:

$$P(B_x) = k\delta\left(\mathbf{s}_x^4, (\text{"}gatc\text{"})\right) \tag{5.1}$$

Where $\delta(a,b)$ is a Kronecker delta like function that is equal to one when $a = b$, otherwise it is equal to zero. k is the binding probability and $\mathbf{s}_a^b$ is the subsequence of $\mathbf{s}$ starting at position a of length b. This simple model is based on the assumption that SeqA binds with equal probability at all locations where the GATC sequence is found.

In this, and all subsequent models, the probability refers to the binding at a single physical location, and will therefore have a value of between 0 and 1.

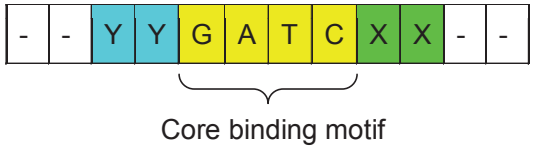


Figure 5-1 Section of genome showing core SeqA consensus motif and flanking dinucleotide sequences. The symmetry of binding is such that the effect of a sequence XX (e.g. AG) will be the same as the effect of the reverse complement at YY (i.e. CT).

The first extension of this model is to add a factor that attempts to model the effect of the pair of adjacent nucleotides on either side of the binding site (Figure 5-1). The model assumes that the core binding probability $P(B_x)$ is modified by a factor which can take a different value for each of the 16 combinations of the adjacent dinucleotide sequence, modelling the degree to which the sequences increases or decreases the probability of binding. This is represented as a four by four matrix **d** where each of the elements $d_{n_1 n_2}$ is the factor associated with the nucleotide sequence $(n_1 n_2)$. The binding site is symmetrical in that the GATC sequence is its own reverse complement. Consequently the effect of a dinucleotide sequence on one side of the binding site is indistinguishable from the effect of the reverse complement dinucleotide sequence on the other. The simplest assumption when modelling is to assume that the effects of the two dinucleotide sequences are independent and multiplicative. The factor for the dinucleotide immediately subsequent to the binding site and also for the reverse complement of the dinucleotide immediately preceding the binding site and these values are used to modify the binding probability within the model.

The extended binding probability can then be represented by:

$$P'(B_x) = kd_{\mathbf{s}_{x+4}\mathbf{s}_{x+5}} d_{\bar{\mathbf{s}}_{x-1}\bar{\mathbf{s}}_{x-2}} \delta(\mathbf{s}_{x,4}, ("gatc")) \tag{5.2}$$

$\mathbf{s}_a$ is the nucleotide at position a and $\bar{\mathbf{s}}_a$ is the complement of the nucleotide at that position. A potential problem with the matrix **d** is that the model fitting is able to modify the coefficients such that this parameter introduces a net gain, and there would therefore be a degree of indeterminacy between the gain introduced by this function and gain introduced by more global gain parameters. This was avoided by introducing a symmetrical mapping between the 16 values of **d** and an alternative 15 value vector (5.3)

$$d'_0 = \frac{1}{16}\left(\sum_{n_1 \in \{a,c\}} d_{n_1 n_2} - \sum_{n_1 \in \{g,t\}} d_{n_1 n_2} \right)$$

$$d'_1 = \frac{1}{8}\left(\sum_{n_1 = a} d_{n_1 n_2} - \sum_{n_1 = c} d_{n_1 n_2} \right) \qquad d'_2 = \frac{1}{8}\left(\sum_{n_1 = g} d_{n_1 n_2} - \sum_{n_1 = t} d_{n_1 n_2} \right)$$

$$d'_3 = \frac{1}{4}\left(\sum_{n_1 = a, n_2 \in \{a,c\}} d_{n_1 n_2} - \sum_{n_1 = a, n_2 \in \{g,t\}} d_{n_1 n_2} \right) \qquad d'_4 = \frac{1}{4}\left(\sum_{n_1 = c, n_2 \in \{a,c\}} d_{n_1 n_2} - \sum_{n_1 = c, n_2 \in \{g,t\}} d_{n_1 n_2} \right)$$

$$d'_5 = \frac{1}{4}\left(\sum_{n_1 = g, n_2 \in \{a,c\}} d_{n_1 n_2} - \sum_{n_1 = g, n_2 \in \{g,t\}} d_{n_1 n_2} \right) \qquad d'_6 = \frac{1}{4}\left(\sum_{n_1 = t, n_2 \in \{a,c\}} d_{n_1 n_2} - \sum_{n_1 = t, n_2 \in \{g,t\}} d_{n_1 n_2} \right)$$

$$d'_7 = \frac{1}{2}\left(d_{aa} - d_{ac}\right) \quad d'_8 = \frac{1}{2}\left(d_{ag} - d_{at}\right) \quad d'_9 = \frac{1}{2}\left(d_{ca} - d_{cc}\right) \quad d'_{10} = \frac{1}{2}\left(d_{cg} - d_{ct}\right)$$

$$d'_{11} = \frac{1}{2}\left(d_{ga} - d_{gc}\right) \quad d'_{12} = \frac{1}{2}\left(d_{gg} - d_{gt}\right) \quad d'_{13} = \frac{1}{2}\left(d_{ta} - d_{tc}\right) \quad d'_{14} = \frac{1}{2}\left(d_{tg} - d_{tt}\right) \quad (5.3)$$

The following are examples of the reverse mappings

$$\begin{aligned} d_{aa} &= 1 + d'_0 + d'_1 + d'_3 + d'_7 \qquad d_{ac} = 1 + d'_0 + d'_1 + d'_3 - d'_7 \\ d_{aa} &= 1 + d'_0 + d'_1 - d'_3 + d'_8 \qquad d_{ac} = 1 + d'_0 + d'_1 - d'_3 - d'_8 \end{aligned} \quad (5.4)$$

The mappings are such that over all of the 16 n_1,n_2 combinations each of the vectors is added and subtracted an equal number of times which ensures that the constraint (5.5) that the average of the 16 values is one is inherent in the mapping.

$$\sum_{n_1 \in \{a,c,g,t\}} \sum_{n_2 \in \{a,c,g,t\}} d_{n_1 n_2} / 16 = 1 \quad (5.5)$$

5.2.4 Modelling of cooperativity in SeqA binding

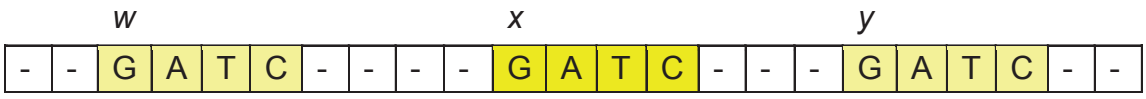


Figure 5-2 GATC motif and two adjacent motifs. The model includes factors for co-operative binding between binding at the site x and the adjacent sites at w and y.

The second extension of the model attempts to incorporate the effect of cooperative binding between nearby binding sites (Figure 5-2). The model assumes that the binding probabilities for two adjacent binding sites are modified by a factor which can take a different value depending on the spacing from one binding site to the next. This is represented by a vector **c** where each of the elements c_a is associated with a spacing a. If the next binding site after the site at x is at location y, and the last binding site before the site at x is at location w then the new probability $P''(B_x)$ that binding will occur at position x has been defined in the model to be:

$$\begin{aligned} P''(B_x) &= P'(B_x) + P'(B_x)P'(B_w)\mathrm{c}_{x-w} + P'(B_x)P'(B_y)\mathrm{c}_{y-x} \\ &= P'(B_x)\times\left(1 + P'(B_w)\mathrm{c}_{x-w} + P'(B_y)\mathrm{c}_{y-x}\right) \end{aligned} \tag{5.6}$$

Again, this equation describes the probability for a single site, and will be multiplied by experiment specific model fitted factors in order to generate numbers that match the experimental results. The factor c_v models the increase (or decrease) in probability of binding at a specific location as a result of the presence of a SeqA protein bound v nucleotides away. This is multiplied by the underlying probability of binding at the other site as determined by the adjacent dinucleotides, on the basis that if the adjacent nucleotides make SeqA binding less likely at a neighbouring site then its potential role in cooperative effects will be proportionately reduced.

5.2.5 Modelling of fragment binding to probe sites

Figure 5-3 represents the process where SeqA being bound to specific locations on the DNA results in DNA fragments being loaded onto probe locations on the microarray.

The fragments can be of the order of 1000 nucleotides in length, and they may have been selected by immunoprecipitation as a result of one (or more) SeqA proteins being bound at any position along the fragment. The fragment length is considerably longer than the distance along the genome between the probe sequences (approximately 100 nucleotides in the case of the microarray used to measure SeqA binding) and so may be bound at any one of the probe positions that overlap the fragment sequence.

The consequence of this is that the probe position to which the DNA fragment is loaded could be up to a fragment length away from the position where the nucleotide was bound.

Figure 5-3 shows the simple one-parameter model that has been used to accumulate the predicted SeqA binding in the region around a probe site in order to predict the fluorescence

intensity F_y seen at the probe position on the microarray corresponding to genomic coordinate y. This can be represented as:

$$F_y = \sum_{x=y-l}^{y+l} P''(B_x)\frac{|y-x|}{l} \tag{5.7}$$

where l is an assumed length of the fragment, whose value is found by model fitting.

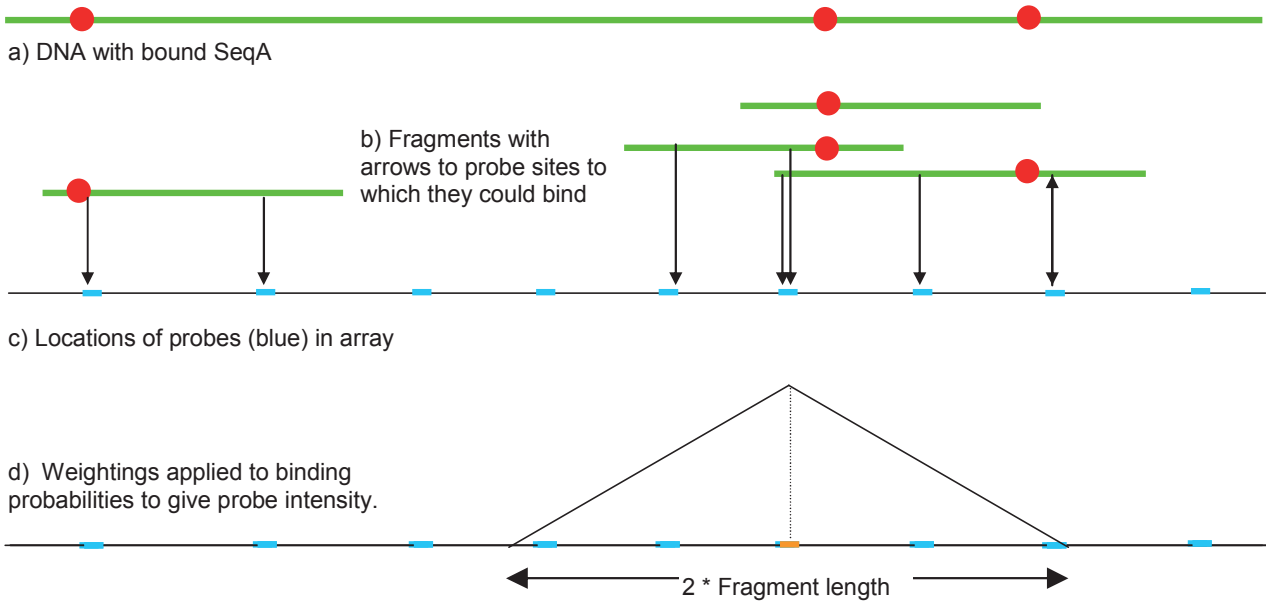


Figure 5-3 Mapping from binding sites to probe sites. a) A section of genome with bound SeqA, which is then fragmented (b). Each fragment spans a number of probe sites which are marked in blue (c) and could bind to any of the probe sites as marked by the arrows. d) shows the weightings applied to the predicted SeqA binding in the region around a probe site in order to calculate the probe intensity predicted by the model

5.2.6 Modelling of global residual parameters

The model so far is based around the probabilities associated with a single binding site. Experimental data will be based on samples which contain many copies of the genome, which is then subject to many additional scaling factors as the sample is processed and sequenced and so the final values from equation (5.7) will need at minimum a global scaling factor to map the model onto the experimental data from an experiment. Initial model fitting showed that there were additional aspects which indicated that are more complex function was required in order to fit the model predictions to the observed data.

The first aspect seemed to correspond to a regional variation in the probability of binding, with SeqA binding across some regions of the genome being slightly more likely than others. The characteristic length of these regions was of the order of 100s of thousands of nucleotides, such that a different scaling factor appeared to be needed for each of these regions. The second was that the model fitting was improved by the addition of an offset to the results predicted by the model. In many ways this is to be expected as there will a number of experimental factors that will result in the zero point for fluorescence measurements being slightly arbitrary. The final aspect was that while the model was successfully able to fit much of the data, it was consistently unable to match the very strong peaks that were seen in the results. A single parameter mapping based on an exponential function was incorporated which added an additional boost to the higher values predicted by the model (see discussion). The final fluorescence intensity predicted by the model is given by:

$$\begin{aligned} F''_y &= \mathrm{g}(y)\exp\left(F'_y + c_2\right)F'_y \\ \text{where } F'_y &= F_y + c_1 \end{aligned} \tag{5.8}$$

g(y) is a gain adjustment that is a function of the coordinate y where the value produced is a constant for each region of 200,000 bases and optimised during the process of model fitting.

5.2.7 Model fitting

Model fitting was performed using an extension of the Amoeba optimisation algorithm [34], based on the Nelder-Mead function minimisation algorithm [74] as this is another model which is non linear and with large number of parameters. The model fitting minimises the sum of squares error function E given by

$$E = \sum_{y=0}^{L}\left(S_y - F''_y\right)^2 \tag{5.9}$$

S_y is the measured fluorescence intensity

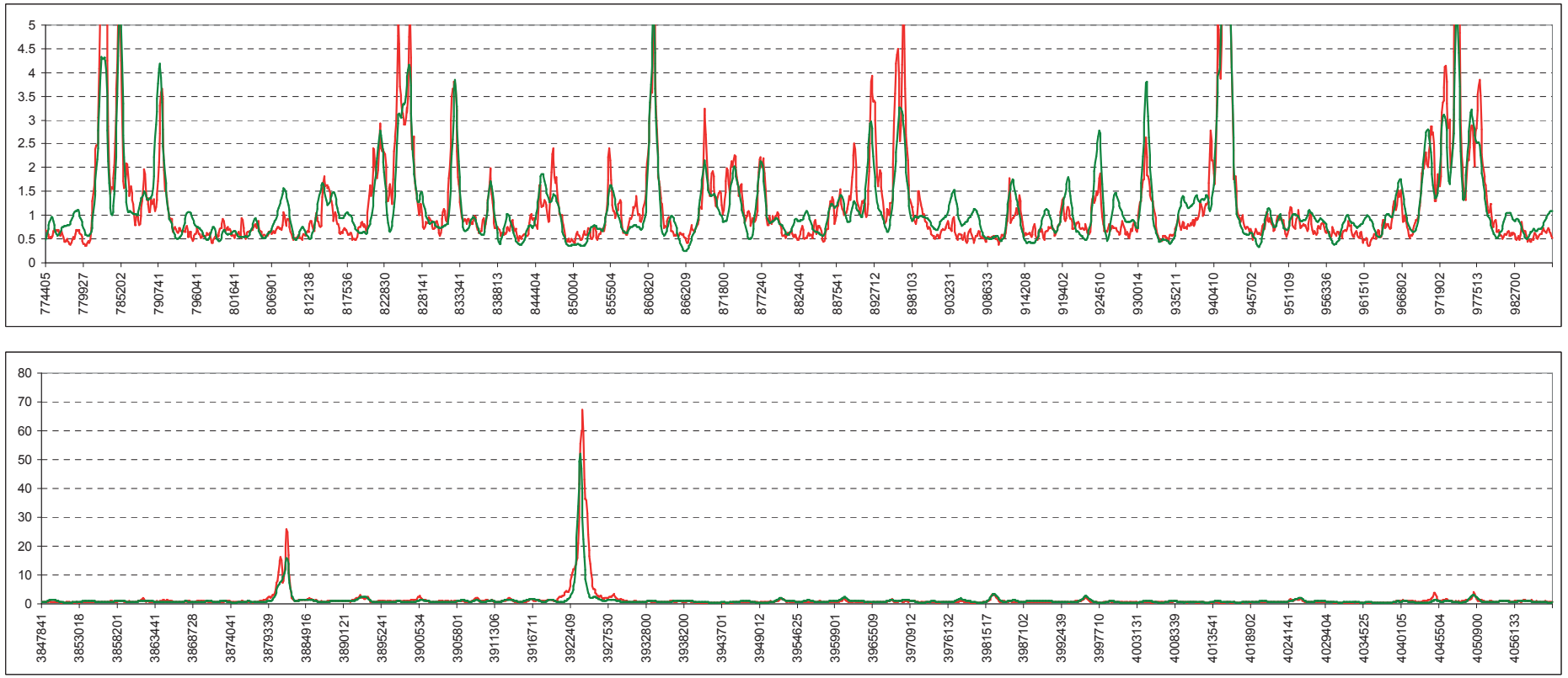


Figure 5-4 Two regions of the *E. coli* genome comparing the observed (red) and model (green) a) shows a region that is representative of the majority of the genome. b) contains the OriC and one other region that is known to have significant levels of SeqA binding.

It was found necessary to add additional terms to the error in order to keep the model fitting well behaved. In particular it was found that some of the weightings for cooperative binding between binding sites were slightly ill defined, possibly because there were insufficient instances of particular inter binding site gaps for the model fitting to be able to distinguish between the effect of different inter binding site gaps. As a result there was a tendency for some parameters to drift in an indeterminate way during model fitting. An additional penalty was introduced that increased with the size of the weighting as follows:

$$E' = E + \sum \left| \mathrm{d}\left(x / 1000 \right) \right| \tag{5.10}$$

This has the effect of causing the parameter to drift towards zero if a non-zero value has no significant effect on the model fit.

5.3 Results

Figure 5-4 shows the degree of match achieved between the observed data and the model after model fitting. It shows two regions, one of which is representative of a large proportion of the *E. coli* genome, and the other which contains the two regions where it is known that there are significant levels of SeqA binding. It can be seen that a good fit is achieved, with the model able to reproduce in considerable detail both the number and relative sizes of the peaks that were reported by the ChIP-seq process.

5.3.1 Regional gain variation

Figure 5-5 shows the variation in regional gains that was identified by model fitting. There is no indication that the model fitting has used the additional freedom that these parameters provide to improve the fit in the region around base pair coordinate 4,000,000 where the major peak in SeqA binding is found at the oriC. The model fitting achieved at this point is due solely to the other model parameters such as the increase in binding predicted as a result of cooperativity between nearby nucleotide sites.

To address concerns that the model fit and the associated parameter values were arrived at partly as a result of over fitting, the process was repeated such that the fitting was done with the data for each half or third genome in turn and the results compared. The initial conditions for the two fitting processes were a set of residual global parameters (Section 5.2.6) that were generated from an initial whole genome model fit.

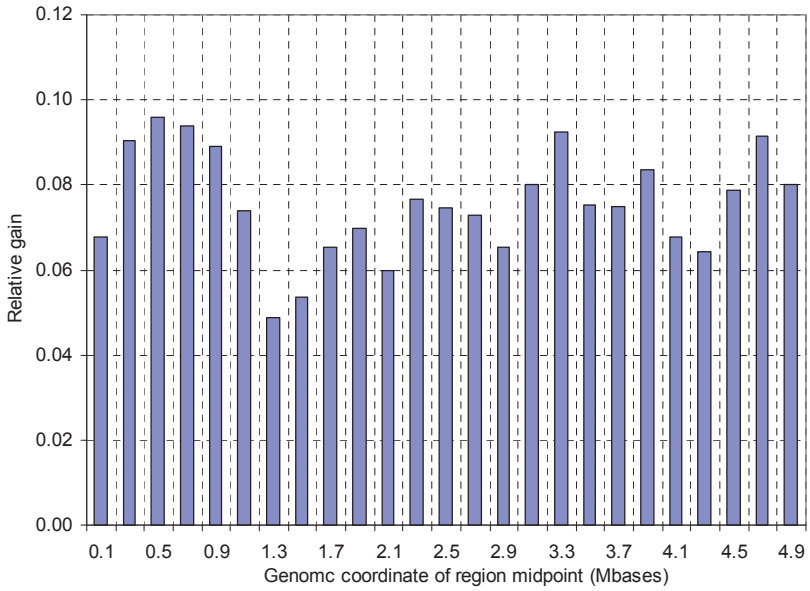


Figure 5-5 Regional gain variation in SeqA binding. The model includes a parameter for each 200,000 nucleotide region of the genome to adjust for apparent regional variation in the degree of SeqA binding. Values shown were generated by a whole-genome model fitting.

5.3.2 Di-nucleotides adjacent to the binding site have a significant effect on binding

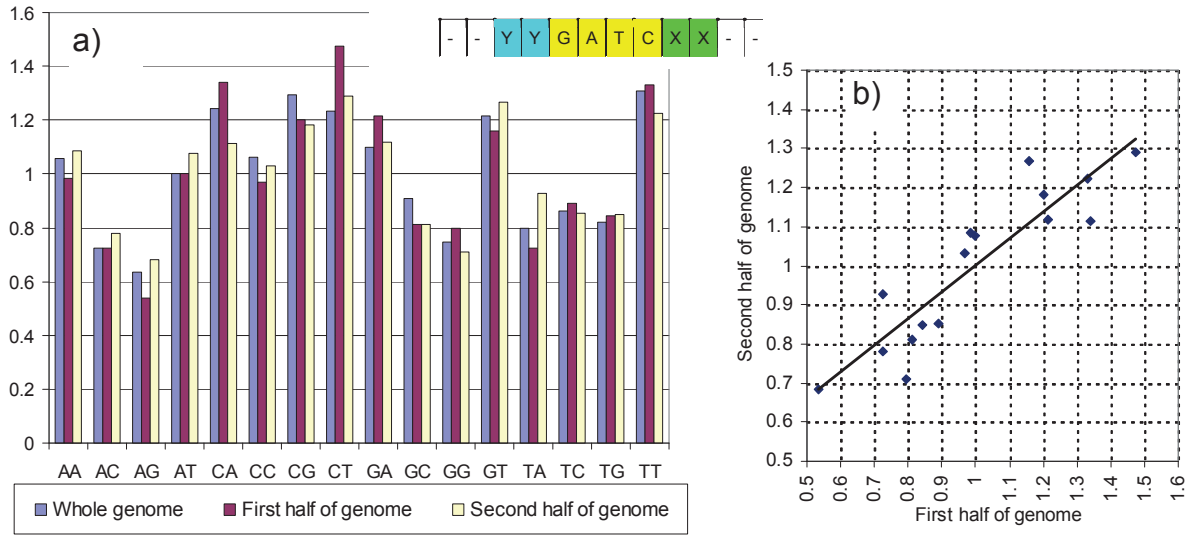
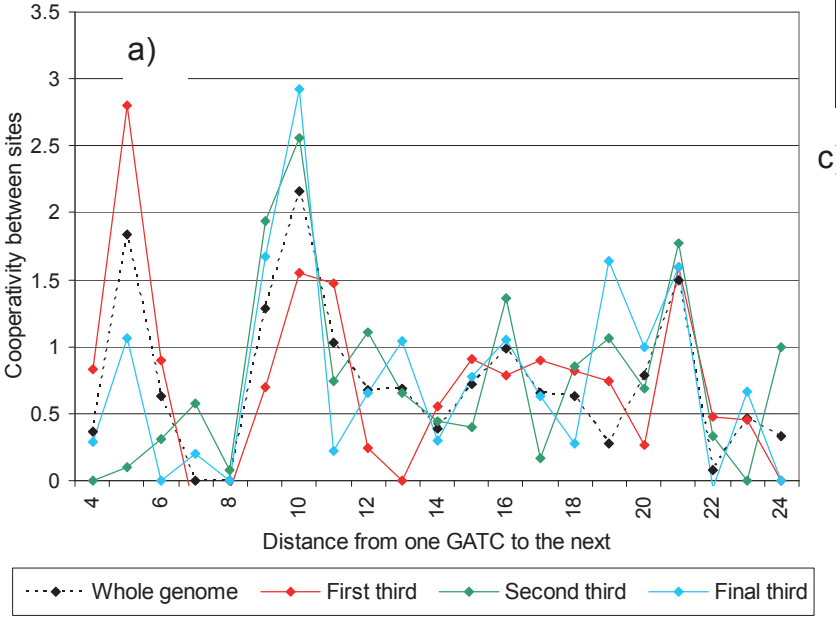


Figure 5-6 Effect of adjacent dinucleotides on SeqA binding. a) The degree to which the adjacent dinucleotide increases or reduces the probability of SeqA binding. The dinucleotides are the pair XX, or the reverse complement of YY in the inset. The values shown are from independent fitting of microarray data from the whole genome and each half genome. b) Correlation of dinucleotide weightings from each half genome. Pearson coefficient = 0.905, p-value = 3×10^{-8}.

Figure 5-6 shows the effect of the adjacent dinucleotide sequence on SeqA binding at GATC locations (Section 5.2.3), comparing the results for the whole genome and the two half

genomes. The good correlation that is seen between the values obtained by fitting to the data from the two halves of the genome is an indication that the values are not as a result of over fitting. The p-value of 3×10^{-8} indicates that the correlation seen between the results from the two half genomes is not consistent with the values obtained from the two half genomes being uncorrelated.

5.3.3 Cooperative effects between adjacent binding sites at specific site spacings



b)		p-value
First	Middle	0.25
First	Final	0.03
Middle	Final	1.2E-5

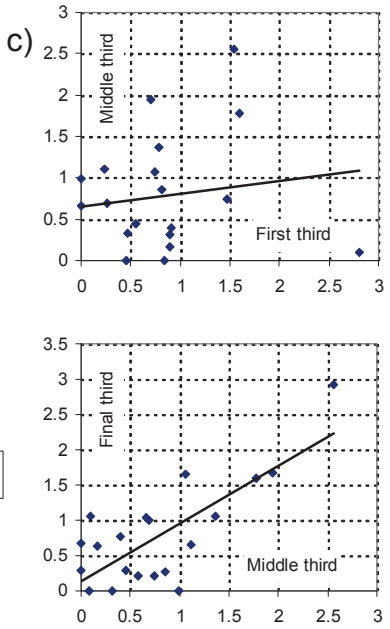


Figure 5-7 Comparison of cooperative binding found by model fitting in each third genome. a) Increase in apparent binding probability as a function of spacing from one GATC site to the next as indicated by the values after model fitting of the vector **c** (Section 5.2.4). A cooperativity of two indicates that the spacing results in additional SeqA binding equivalent to there being two extra sites at each location. Values were determined independently from data for each third of the genome, and also for the whole genome. b) p-values for the correlation between the data from each 1/3 genome. c) x-y plots showing the correlation between the results from the first and middle third, and also from the middle to the final third.

When the results for cooperative binding for GATC spacing of up to 25 nucleotides were calculated for the two half genomes and compared, the Pearson coefficient was 0.388, and the p-value for significance of the results was 0.04, only just significant at a 5%

confidence level (data not shown). In order to understand this a little more the process was repeated dividing the genome into thirds (Figure 5-7).

This suggests that there is clear information coming from the model fitting exercise about the cooperative or lack of cooperative effect between binding sites at certain spacings, particularly spacings in the region 7 to 10 nucleotides, and also 21 nucleotides. Data for other spacings appears less well defined although it does suggest a degree of cooperativity exists when the spacing between binding sites is between 10 and 20 nucleotides.

5.4 Discussion

These results can be considered from two angles: first there are general observations about the use of data modelling techniques for the analysis of ChIP-chip data and second, the interpretation of specific data on SeqA binding.

5.4.1 The application of model fitting to interpreting ChIP-chip data

The results of analysing this one dataset suggest that model fitting could be a very useful tool in the bioinformatician's armoury for extracting information from raw ChIP-chip data. Rather than concentrate on locations where there is a strong signal, this technique makes use of information from all of the probe sites, and with this additional information comes the potential to extract more subtle information about the characteristic in question using the microarray.

One lesson drawn from this initial study is the importance of validating the model parameters which are obtained. Cross-validation techniques are particularly useful, and have proven particularly effective in identifying the significance of the role of adjacent dinucleotides and cooperative binding in this particular case. Cross-validation also showed that the results taken as a whole for the cooperative effect between proximal binding sites based on the modelling technique were only just significant (p-value = 0.04). Within the results it appears that the results for certain binding site spacings are more significant. Such results can be used to inform the process of modifying and improving the model, as was the case during earlier stages in the investigation when the initial results were used to suggest improvements that could be made to the model.

Cross validation showed that the results relating to the effect of the adjacent pair of nucleotides were particularly significant and robust (p-value = 3×10^{-8}), suggesting that this may be a particularly useful tool for getting more detailed information about binding site

preferences from ChIP-chip data. This was confirmed in a subsequent brief study looking at MntR binding in *E. coli* (results unpublished).

Modelling can also provide a better understanding of how the general pattern of ChIP-chip results arise. Many of the peaks in Figure 5-4 are the result of a small cluster of closely spaced GATC binding sites, and the width of the peak is a result of the effect of the fragment length causing fragment binding over multiple probe sites. This is catered for in the model, which indentified an average fragment length of 1040 as being consistent with the results. The fitting of this parameter will be driven in part by the fitting of the shapes of these peaks.

The model then predicts the extent of the effect of such a fragment length on the smoothing of the results. The actual fluorescence measurements show significantly more variation between adjacent probe sites than would be predicted by the model, which gives a measure of the degree of noise that is introduced by the random variation in fragment binding and the uncertainty in probe fluorescence measurement.

The results show the degree to which the information from individual binding sites is distributed over a significant number of probe sites as a result of the fragment length being significantly longer than the inter-probe distance. This suggests that even if only conventional techniques were being used to identify SeqA binding regions, they could be identified with more accuracy if the average fragment length was shorter.

There are, however, restrictions and limitations to this technique, some of which were highlighted in the SeqA data which were examined. The technique relies on the contributions of other effects being minimal or identifiable. In the case of the SeqA data one such characteristic was an apparent regional variation in SeqA binding affinity across the genome. It has not been possible to identify the cause of this effect, but it was possible to incorporate a model component that will correct for this effect.

5.4.2 Adjacent dinucleotide sequence has a significant effect on SeqA binding in *E. coli*

Perhaps the most significant result of this investigation is the information about the role of the adjacent pair of nucleotides in determining the probability of SeqA binding at a specific site (Figure 5-6).

The model assumes that the contribution of the dinucleotides on either side of the binding site is multiplicative, and independent. This means that SeqA is four times more likely to bind to the sequence AAGATCTT than to the sequence CTGATCAG. This effect has not been identified in any previous experimental work involving SeqA and *E. coli.*

5.4.3 Cooperativity in the binding of SeqA to *E. coli*

Cross-validation of the results predicting cooperativity of SeqA binding from ChIP-chip data suggest that, with a p-value of 0.04, the overall results only just warrant rejecting the null hypothesis assuming the standard threshold of a 5% significance level. This means that conclusions should be drawn from these data with some caution.

One general conclusion from model fitting is that an improved fit was obtained as a result of adding cooperativity between sites that were spaced up to 24 nucleotides apart (Figure 5-7). For some of the spacings (e.g. 11 to 15 and 17 to 19) there is a degree of indeterminacy in the degree of cooperativity predicted by the model, suggesting that there is a degree of over-fitting taking place. The resulting lack of correlation results in the poor p-value that is seen for these results.

There are some locations where model fitting indicates that there is a consistent lack of cooperativity, particularly for GATC spacings of eight nucleotides. There are other spacings, most notably 10 and 21 for which there is consistent evidence of cooperativity. It has been suggested that there is also some degree of cooperativity between sites that are spaced 16 nucleotides apart. The results for five nucleotide spacing are a little difficult to interpret, with a strong indication of cooperativity present in data from the first third of the genome, and virtually no evidence from the second third.

These general observations are consistent with previously published results derived from the same data (Figure G-1). These showed that GATC spacings of 10 and 21 nucleotides are more frequently found in regions surrounding SeqA binding peaks. The previous results also show that spacings of 19 nucleotides are found, and there is a slight suggestion of this from the model fitting data as well.

It is already well established that SeqA binds cooperatively to DNA and extensive information has previously been obtained on the effect of the spacing between binding sites on the probability of binding *in vitro*, using artificially constructed nucleotides [15]. These results also show a preference for binding when GATC sites are spaced in multiples of approximately 10 and 21 nucleotides (Figure 5-8).

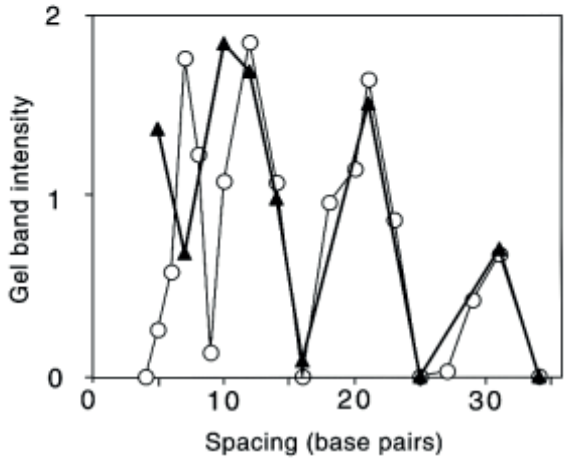


Figure 5-8 Effect of GATC site spacing on GATC binding obtained using constructed oligonucleotides [15]. This shows the relative SeqA binding on two hemimethylated GATC sites at various spacings on short lengths of artificial double stranded DNA. The open circles indicate methylated sites which are on the same strand, and the triangles indicate sites on opposite strands.

Chapter 6

Conclusions and further work

The conclusions and further work that arise out of the previous sections are divided into a number of different categories and considered separately. The first relates to the use of modelling techniques to draw more information out of microarray and high throughput sequencing data. The next categories relate to sequencing bias results that were obtained using this technique and the information about protein binding in prokaryotes. The final category is the technique for obtaining more information about protein binding from ChIP-seq data.

6.1 The use of modelling to extract additional information from genomic data

6.1.1 Conclusions

The sequence bias results in Chapters 2 and 3 and also the binding site information in Chapter 5 shows that the model fitting technique that has been used is a very powerful tool that can be added to the armoury of tools that is available to the bioinformatician. In some respects this can be viewed as a combination of two existing techniques that have a long pedigree.

The first is the principle of ensemble averaging where data is combined from multiple locations in the genome such that information that common to each location will be correlated and will combine additively, whereas uncorrelated random data from each location will not. This increases the signal to noise ratio of the information associated with the sequence by a factor of the square root of the number of locations used [13]. This principle is used to combine fragment distribution information for each of the 8-mers in the genome in chapter 2 where it was shown that the values of $Y_{\mathbf{s}}$ had a Poisson distribution with a standard deviation was $1/\sqrt{E_{\mathbf{s}}}$ and so the uncertainty or noise associated with the calculated value reduced in proportion to the square root of the number of samples used. This principle also underpinned the use of fragment information relating to the incidence of each potential binding motif in Chapters 4 and 5.

The second technique is the use of modelling and model fitting in order to identify the underlying common characteristics of the feature being investigated. This method is used to identify the relationship between sequence and fragment start positions in Chapters 2 and 3

and also further details of the relationship between sequence and the degree of protein binding in chapter 5.

The power of these techniques when used in this way is that they are able to make much more use of the data that are available. This is partly because this technique uses both positive and negative information, e.g. information about the 8-mers which increase the probability of a fragment start and the 8-mers which reduce the probability of a fragment start. While techniques such as MEME also do this, in that they use information about regions where a motif is present, and the background distribution where it is absent, they do so in a much less powerful way. This is because the information is reduced to a binary yes/no value, e.g. dividing the genome into 'peak' and 'not peak'.

The modelling technique uses the full analogue measurement data such as peak height or ChIP-seq binding strength, allowing subtler aspects of the property under investigation to be drawn out through modelling.

6.1.2 Further work

The description of the Nelder-Mead model fitting algorithm suggests that it is well suited to model fitting applications such as the ones described in this thesis. Potential problems of any model fitting algorithm include non-optimal speeds of convergence, failing to find a global minimum and getting stuck at a local minimum. During the investigation the algorithm appeared to perform well in both regards, and consequently no time was spent investigating other alternative parameter optimisation algorithms. In addition this algorithm is still used at the heart of other optimiser software that is widely used such as EcosimPro [66].

It is nevertheless the case that it would be worth investigating alternative algorithms, especially in view of the extensive work in this area since the Nelder-Mead algorithm was published, to see if there might be other more recent algorithms that might give better performance.

One possible algorithm that may give improved performance is the Direction set or Powell's minimisation algorithm [81]. As with the Nelder-Mead algorithm it does not rely on the ability to calculate gradients and also uses a set of unit vectors. Its approach however is to generate a set of non-interfering or conjugate vectors so that minimisation can proceed by minimising with respect to each vector in turn.

6.2 Sequence bias in next generation sequencing data

6.2.1 Conclusions

The main conclusion is that the pattern of DNA and RNA fragmentation that occurs during protocols such as ChIP-seq and RNA-seq is considerably more complex than has previously been recognised, and that this can provide information both on the process of fragmentation that occurs during the procedure and also on the proteins that are bound to the DNA.

Particularly significant is the finding that in both ChIP-seq and RNA-seq there are multiple alternative patterns of sequence bias that coexist within the same experiment, and that there are significant differences between experiments. While this information can be used to correct for the bias, the results also suggest that any correction would not significantly affect the existing use of such data to identify binding sites, gene expression levels and the usage of alternative transcripts. They appear to be important when attempting to extract subtler information from the ChIP-seq data, such as fragment distribution fingerprints in regions associated with over-represented motifs that were described in chapter 4. It is possible that there may be other equally subtle information that can be extracted from ChIP-seq data after they have been corrected for sequence bias.

6.2.2 Further work

This work has led to the generation of a number of hypotheses about what happens at the molecular level during the process of DNA fragmentation. These include:

- The existence of various mechanisms with different characteristics that occur to varying degrees in different experiments (2.5.1-2.5.6)
- The potential role of double stranded DNA coming together to form quadruplexes with other DNA and catalyse the creation of further fragments which share similar end sequences (2.5.4).
- The relationship between sequences that are more likely to break during sonication and the nucleotide distribution in the genome (2.5.7).

Further experimental work, such as deliberately introducing strands with specific sequences in order to artificially bias the DNA fragmentation is required to verify these hypotheses. For example, this more detailed understanding of the processes that occur during the ChIP-seq protocol may create the possibility of deliberately biasing the fragmentation process to obtain more information about specific regions or sequences when performing

ChIP-seq experiments. Such biasing could also be used to influence the fragment distribution in experiments using *Arabidopsis* in order to increase the number of fragments created in the promoter regions and improve the quality of the data compared to that which is currently available.

6.3 Obtaining information about protein binding from ChIP-chip data

6.3.1 Conclusions

The work with SeqA data from *E. coli* shows that the modelling technique that has been developed allows significantly more information to be obtained about protein binding in prokaryotes.

6.3.2 Further work

This thesis describes the successful application of this technique to one prokaryotic transcription factor. This opens up the possibility of repeating a similar analysis on other data, a process that has already started with the analysis of some ChIP-Selex data relating to MnTR binding (unpublished).

Some of the more subtle information that has been found could have been obtained using ChIP-seq, but it is also the case that applying the same techniques to ChIP-seq data could allow even more information to be obtained from the raw data.

One of the reasons behind the success of this technique when applied to prokaryotic data is that the pattern of binding is much simpler, with proteins able to bind to a roughly equal extent at all locations with identical local sequences. In the case of eukaryotes the landscape is much more complex, with factors such as the chromatin structure making some regions inaccessible, so that proteins will bind to very different extents at two locations with identical local sequences.

It may be possible to develop the model so that it takes such factors into account. In this way, it would be possible to extract more detailed information about eukaryotic binding sites.

6.4 Obtaining information about protein binding from ChIP-seq data

6.4.1 Conclusions

Although this was the starting point for a great deal of this work, the time spent investigating the other factors that influence DNA and RNA fragmentation means that the study of this effect is still at an early stage. The initial results do however confirm the original

belief that ChIP-seq data are able to provide more information about the binding of proteins to DNA than is currently made available using existing techniques for analysing ChIP-seq data. In particular, knowledge about how the presence of protein on DNA influences the probability of DNA fragmenting during sonication may provide additional information about the nature of the binding between DNA and proteins.

6.4.2 Further work

Considerably more work, perhaps combining this technique with the modelling techniques that were applied to the ChIP-chip data, needs to be carried out in order to understand the full extent of the information that can be obtained from ChIP-seq data using this technique.

One factor that needs to be considered is that, at the point of sonication, the protein-DNA bond is modified by the presence of formaldehyde which keeps the protein bound to the DNA. The significance of this with regard to the results that have been obtained needs to be explored further.

The work thus far has shown that correcting for global sequence bias makes a significant difference to the apparent pattern of fragmentation in the region of a protein-binding motif. More work is required in order to determine how appropriate such corrections are in the immediate vicinity of a bound protein, as the mechanisms that cause the effect that is seen globally may not apply to a region where a protein is bound.

The tentative conclusion that the NRSF fingerprint may provide information on its action as a chromatin remodeler (Section 4.4.4) would appear to be a good example of the information that could be made available using this approach. Further work would be required to confirm that this is an indication of chromatin remodelling. If so, then the size of the peak may be sufficient to allow the ChIP-seq data to be used to determine the extent of chromatin remodelling at individual sites. Initial results indicate that the extent of the remodelling is related to the match between the sequence and the binding motif. The ChIP-seq data may be able to provide more detail as to which regions of the extended motif are more important in order for chromatin remodelling to occur.

Appendix A

A Method for locating non-unique regions in genomes

A-1 Introduction

Background: The problem of separating different mechanisms in model fitting

The motivation for the research described in this thesis is to determine whether there is additional, biologically informative, information that can be derived from the non-random distribution of fragments that occur when DNA and RNA is fragmented in high throughput sequencing protocols. The challenge in any such investigation is to separate out the different mechanisms that give rise to the non-random distribution that is seen. There are two reasons for wanting to separate such mechanisms, which relate to the degree to which the variation introduced by a mechanism is correlated to the variation introduced by the other mechanisms.

If the variations introduced by different mechanisms are essentially uncorrelated then the variation introduced by one mechanism can be considered as being noise as far as the other mechanisms are concerned. The presence of any noise will always have a detrimental effect on the use of data to characterise the mechanism that generated the data, in that it determines the precision of any results that are derived from the data. Any process that reduces the noise will improve the way that the data can be used to characterise the mechanism.

If the variation is correlated to any extent then it will more difficult to distinguish between the effects of the two mechanisms. In such as situation the only option available is to see whether some other approach can be used to separate the two sources of variance, making use of some characteristic of the experimental design and analysis that allows the source of variation to be removed.

The problem of non-unique regions

Any investigation into the way that the nucleotide sequence influences the DNA fragmentation must take into account the problem of sequences which can be aligned to multiple locations in the genome. A common convention is to discard such sequences rather than introduce errors as a result of making assumptions about where the sequences originated. The regions of the genome where fragments cannot be aligned in this way are often referred to as unmappable. In order to locate where the DNA or RNA fragments originated in the

genome, the first N nucleotides of each fragment are sequenced and then aligned to the genome by programs such as bowtie [57] or SOAP [62]. Even if the genome sequence had characteristics equivalent to a fully random sequence, at any given read length there would be a certain proportion of sequences which occur more than once in the genome, and so it is not possible to identify such sequences with a unique location on the genome. The regions in the genome where these sequences are found are consequently unmappable at the given read length. In practise, genome sequences are not random, for example there are significant proportions of the eukaryotic genomes that consist of relatively repetitive sequences, and this increases the proportion of the genome that is unmappable.

The approach that is frequently adopted, including in the work described in this thesis, is to exclude or ignore all sequences that cannot be uniquely mapped to the genome, which is done, for example, using the '-m 1' option in Bowtie. The result of this decision is that one of the reasons why the fragment distribution appears non-random in some regions is not because there are no fragments in the region but because the region is unmappable.

As a toy example, consider a hypothetical case where the vast majority of 25 nucleotide sequences beginning with the sequence ATGGATGG are unmappable sequences. Consequently there would be very few sequences beginning ATGGATGG in the aligned ChIP-seq data despite there being many instances in the genome. A superficial analysis that ignored mappability would conclude that fragments that start with the sequence ATGGATGG are underrepresented. In an analysis of the effect of genomic sequence on fragmentation the incorrect conclusion that the sequence ATGGATGG somehow suppresses fragmentation could be drawn.

This example demonstrates how there is the potential for correlation between the effects of mappability and the potential effects of nucleotide sequence on fragment distribution, and therefore the importance of correcting for the effect of mappability before making any further analysis of the data.

Creating a map of non-unique regions

In order to compensate for the effect of unmappability, it is necessary to make a map of all the non-unique regions in the genome being analysed.

At the time of the start of these studies there were no publically available tools that had been created to identify these regions; tools such as GEM mapper [24] which was used to generate the mappability tracks in the ENCODE genome browser [84] only became available in early 2010. Based on published Bowtie performance measurements [57] it was estimated

that such an approach would take approximately 120 hours, but would also require significant development to create the appropriate input files, and analyse the output files in order to create a mappability file in a suitable format. In addition, the intermediate file sizes would be considerable as Bowtie was not engineered to perform such an analysis in a file efficient manner and the transferring and processing of such files could in itself add considerable extra time to the core processing time.

Consequently the decision was taken to create a software tool to create these maps. One additional advantage of creating such a tool is the possibility of using the design as a basis for performing other, similar analyses of the DNA sequences.

The approach adopted was to create a hash table of the complete genome sequence in a form that was optimised for finding the maximal length of the region of DNA elsewhere in the genome that aligns with the target DNA sequence, whilst at the same time minimising the memory footprint of the index. The approach did not employ techniques such as the Burrows-Wheeler [17] transform to reduce the hash table size as used by alignment algorithms such as Bowtie. This transform works by converting the genome sequence into a more compressed format prior to creating the hash table.

One of the other problems with sequence alignment is how to handle sequences which do not map at all, but can be mapped if allowance is made for one or two mismatches. Such mismatches may arise as a result of read errors, or may arise as a result of the presence of Single Nucleotide Polymorphisms (SNPs) within the sequence. However, this does not affect the locations of regions that are unmappable in the genome. A sequence that can only be mapped to the genome if mismatches are allowed for will still not map to an unmappable region because, after mismatches, it would align to multiple locations and be rejected. In addition, allowance for mismatches does not create additional unmappable regions. For example, if a sequence maps to one location with no mismatches and another with mismatches, the no-mismatch mapping takes precedence and the sequence is mapped, which means that the first region does not in some way become unmappable because of the existence of the presence of similar unmappable regions.

A-2 Method

The process of producing a listing of unmappable regions consists of three phases. The first creates a form of hash table of the genome. This phase is split into two stages in order to keep a manageable memory footprint, with intermediate results being written to file. This hash table can be used for calculating mappability for any sequence size above a threshold.

The second phase produces a semi consolidated list of mappings between regions of the genome with identical sequences, ordered by the order in the genome of the A sequences. The output of this phase forms a useful source of data about the way in which any specific region might be unmappable. The final phase is then to process the list to form a consolidate table of the regions in the genome that are unmappable.

The hash table design

The simplest hash table is an index that identifies the location of every instance of every possible N-mer in the genome, with the length of the index being determined by the length of the N-mer; increasing the value of N increases the number of indexes, but reduces the average length of the indexes. The value of N is also determined by the combination of the minimum sequence length that will be aligned to the genome, the maximum number of mismatches allowed and details of the search algorithm that uses the indexes. Although the mismatch feature is not required when making creating maps of non-unique sequences the search algorithm was designed to cater for up to two mismatches so that it could be used in other applications. The search algorithm was also designed to enable non-unique sequences with lengths down to 25 nucleotides to be mapped. These requirements meant that the indexes need to identify the location of 11-mers in the genome, although the algorithm was designed so that an entry was only required for every 9^{th} position. For each match, a record is then kept of the sequence of eight nucleotides on either side of the 11-mere. There are therefore $4^{11} = 4194304$ hash tables, and for the human genome each contains a mean of 81 entries in each table. In practise the figure of 81 will vary significantly with sequence, with there being significantly more than 81 for sequences that are associated with repetitive regions.

These tables are only required for the DNA sequence in the forward direction.

Figure A-1 shows how the hash table is then used to identify the location of a sequence in the genome. This could be in order to locate a sequence tag, or the sequence could be an extract from the genome in order to determine whether this sequence has a match elsewhere in the genome and is therefore non-unique. The search algorithm can locate one or more locations in the genome that match a sequence, and can track each of the matches to see which extends furthest. The algorithm has been designed to allow for efficient identification of sequences that match for considerable distances, as is frequently the case when searching for non-unique sequences, without having to store a complete copy of the genomic sequence as well as the data required for the indexes.

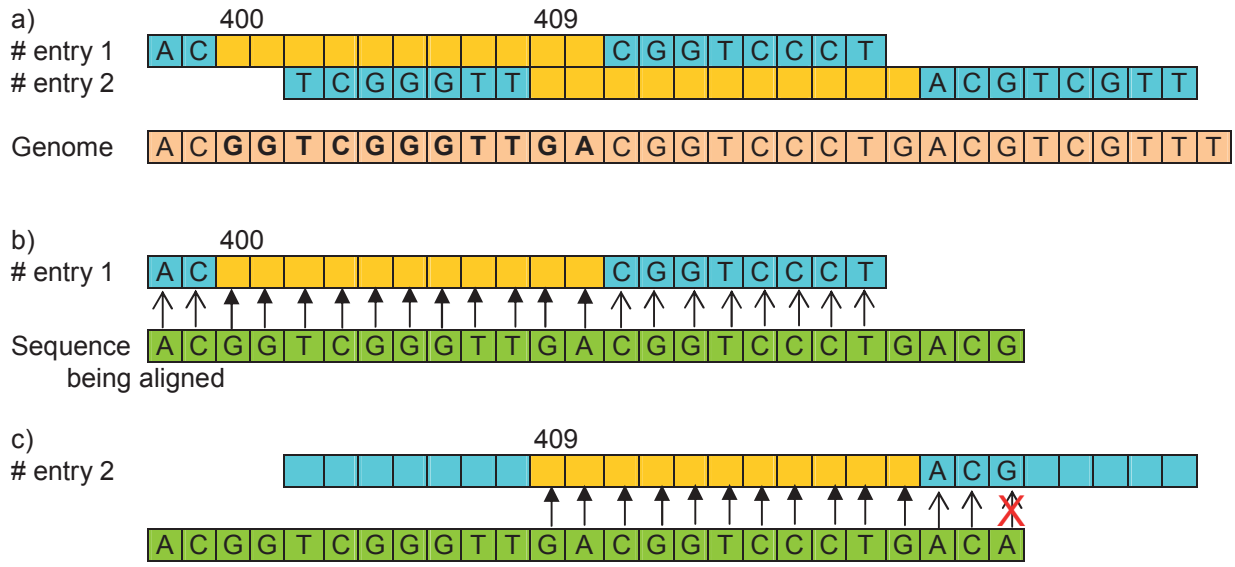


Figure A-1 Sequence alignment using the hash table. a) Two hash table entries in different indexes and their relationship to the genomic sequence. # entry 1 is in the index of all the locations of the 11-mer GGTCGGGTTGA (orange) and indicates an instance starting at position 400 in a chromosome. It also contains information about the adjacent eight nucleotides (blue). Also shown is a second entry in a different index identifying the equivalent information for the 11-mer GACGGTCCCTG starting at position 409. b) When aligning a sequence (which in this example matches the genomic sequence in a) (green) an 11-mer (GG..GA in the example) is selected from the sequence, and all of the entries in the associated index are searched to identify one or more entries where the nucleotides on either side also match. A match is shown for the entry for location 400 in the chromosome. c) For each match, the index for the 11-mer nine nucleotides along is searched to see of there is an entry that corresponds to a position nine nucleotides along from the matches found in the first index. For each match the information relating to the adjacent nucleotides is searched to determine how far the match continues.

If a match is found that is longer than the sequence length S being considered (e.g. 25) then an unmappable region has been located. The process of looking for further unmappable regions can then continue with a seed that starts S nucleotides before the end of the region that matches a sequence elsewhere in the genome.

If no match is found then the process repeats with the seed sequence starts one nucleotide to the right of the previous sequence, and continues in this way until a match is found. The hash table allows maps of mappable regions to be created with sequence lengths down to 19 nucleotides. Below this length the use of entries for one in nine locations can mean that there may not be a hash table entry that provides the required link between the search and target sequence.

Hash table memory footprint

Table A-1 shows the data requirements for each entry in the hash table, showing that a total of 72 bits are required. An entry is required for every nine nucleotides which would take 18 bits to encode using a naïve encoding algorithm with no additional compression. This represents a data expansion of a factor of four, and so the memory footprint for the core data in the hash table for the human genome is approximately 3 GBytes. The implementation of the algorithm gave a total program memory footprint of just under 8 GBytes, the additional memory arising from the overheads of storing the data tables in memory. While not as efficient as Bowtie, this is a not unreasonable requirement for current computing platforms.

Data type	Range requirement	Data stored as:	Number of bits
Chromosome identifier	0-256	unsigned char	8
Location in Chromosome	$0\text{-}2^{32}$	unsigned 32 bit integer	32
Preceding 8 nucleotide sequence	8 * 2 bits	unsigned 16 bit integer	16
Subsequent 8 nucleotide sequence	8* 2 bits	unsigned 16 bit integer	16
Total			**72**

Table A-1 Data requirements for each entry in the hash table

Phase 1: Creation of the hash table.

The first phase of operation is to step through the genome in nine nucleotide steps, creating the hash table entries. These are sorted into 128 different groups, depending on the first nine bits of the 22 bit sequence identifier. The entries are held in memory until a threshold is reached for a specific entry when the results for that entry are transferred to a file, one for each group. Once this has been completed, each of the files are then processed, bringing together all of the entries associated with a single sequence, and then storing the file in a binary format together with an index file that points to the start of the hash table for each of the sequences contained in the file.

This data can then be used to create an unmappability map for any sequence length of 19 nucleotides or greater.

Phase 2: Identification of unmappable segments for a given sequence length

The second and third phases must be repeated for each sequence length S for which unmappability data is required. In the second phase the complete genome is processed

nucleotide by nucleotide, and for each position, the hash table index for the 11-mer sequence starting at this position is used to search for matching sequences elsewhere in the genome.

At each position the match search algorithm looks backwards for up to eight nucleotides as well as forward for as far forward as the match is found to continue. The algorithm monitors the most recent match that extends furthest to the left (M_L) and the most recent match that extends furthest to the right (M_R) with respect to the direction of the forward strand. When a new match is found that exceeds N then the following is done:

- If the new match extends either of these matches in both directions then the previous match is replaced with the new match.
- If the new match lies within either of the matches then the new match is discarded
- If the new match extends further to the right than M_R then M_R is logged to a file and replaced by the new match information
- If the new match extends further to the left than M_L then M_L is logged to a file and replaced by the new match information

This process is performed in two stages. In the first stage the whole genome is searched looking for matches between the forward strand and the forward strand, and in the second stage, the search is for matches between the reverse strand and the forward strand data. The second stage steps through the reverse complement of the forward strand looking for matches to the forward strand. The match information obtained is adjusted and recorded to file as a match between the forward strand and the reverse complement strand.

The output from Phase 2 is two files, one containing forward to forward strand mappings and the other forward to reverse strand mappings. Each mapping entry also identifies the matching region elsewhere in the genome. There will be considerable overlap within this information, e.g. there may be an entry indicating that nucleotides 100 to 150 map to one region and another indicating that 120 to 190 map to another region.

Phase 3: Consolidation of mappings

The data in both of the files from Phase 2 are then consolidated to produce a single file that contains information about the regions where the sequence is non-unique. The two overlapping regions previously mentioned would be considered as indicating that the nucleotides in the region from nucleotides 100 to 190 all map to regions elsewhere is the genome with lengths of greater than N nucleotides.

This is then converted to a file that indicates the regions in the genome where it is not possible to map a sequence starting at that location to a unique location in the genome. If the

sequence length is 25 and there is a region 25 nucleotides long that aligns to a region elsewhere in the genome then this results in a single location in the genome as being unmappable. In the previous example, the region from 120 to 190 results in fragment unmappability from nucleotides 120 to 166.

A-3 Results

The non-unique region mapper has been used to create unmappability data for a number of different genomes for a number of different read lengths. For the human genome the generation of the original indexes takes of the order of 40 minutes on a 3GHz Intel Xeon 5160 processor using one of the cores. The creation of a mappability file for a given sequence length takes about 50 hours cpu time on the same processor.

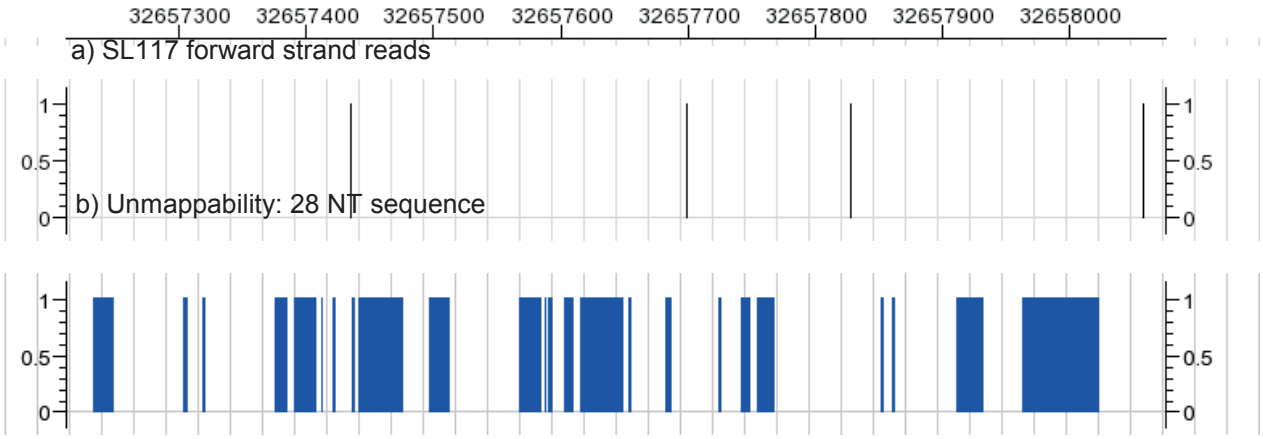


Figure A-2 Comparison of published ChIP-seq data and unmappability. a) Distribution of fragments starts in SL117 dataset with 25 nt read length b) Unmappable regions for 28 NT sequence

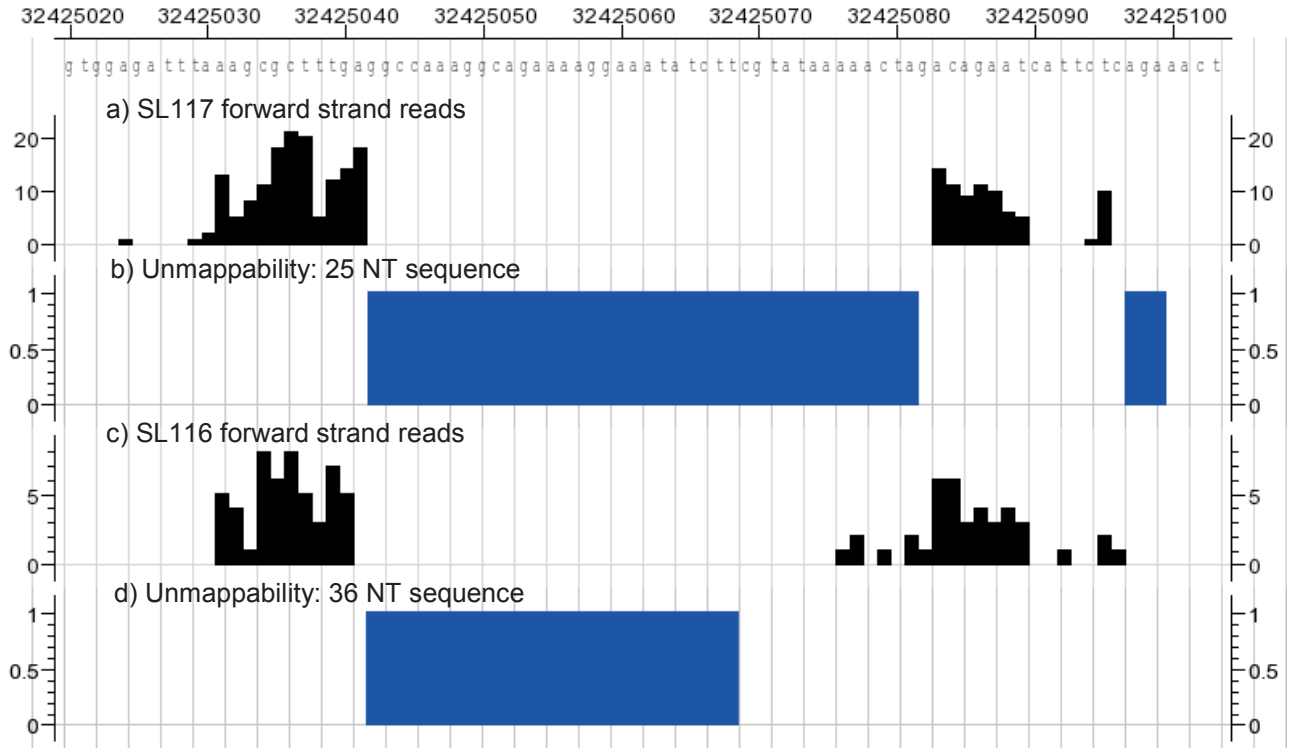


Figure A-3 Comparison of published ChIP-seq data and unmappability. Nucleotides 32425020 to 32425103 of Chromosome 19. a) Distribution of fragments starts in SL117 dataset with 25 nt read length b) Unmappable regions for 25 NT sequence. c) Fragment distribution for SL116 with 36 nt read length and d) Unmappable region for 36 nt sequence.

The cisGenome software has been extended to allow the display of files containing region information (Appendix G-2). This allows the display of the unmappable region data and the comparison with the fragment distribution of ChIP-seq data.

Figure A-2 shows a fairly typical region with widely distributed fragments, where the distribution is consistent with fragments not starting within regions marked as unmappable

Figure A-3 shows a region where an artefact in the process results in clusters of sequence tags that are bunched together. The two datasets were sequenced to different read depths, and in both cases it can be seen that the tags do not extend into the regions indentified as unmappable, demonstrating the consistency between the mapping process performed by the Myers/HudsonAlpha lab and the identification of unmappable regions.

A-4 Discussion, conclusions and further work

Although subsequently overtaken by the availability of other public domain software such as GEM mapper [24] the development of the unmappability mapper as part of this research made a vital contribution to the analysis performed in the research documented here,

and also provided a source of mapping information for other bioinformatics research at the University of Warwick. Its performance is comparable with other mapping algorithms, so there as been no cause to switch to using alternative software.

The intermediate mapping data generated during the process is potentially of value for investigating genetic sequence similarity characteristics, although this has not been pursued.

The algorithm was designed to be able to perform sequence alignment with up to two mismatches, and this has been validated with one set of sequence tag data. The performance achieved appeared comparable to that of many publically available aligners. This development has not been pursued however as it did not appear to offer any advantages over these publically available aligners.

Appendix B

Additional nucleotide bias results ‡

Section 2.2.4 introduced an information content based approach for determining the degree to which each nucleotide in the region proximal to a fragment start site influenced the likelihood of seeing a fragment start at that position. The average of the set of values associated with the nucleotide position indicates the degree to which the nucleotide at that position influences the likelihood. The standard deviation indicates the degree to which this is affected by the neighbouring nucleotides.

The following figures show the average and standard deviation of this value for a range of different sets of ChIP-seq input data. Positive nucleotide values indicate nucleotide positions within the fragment and negative values indicate the nucleotides immediately preceding the fragment.

In some cases the 8-mer either covers the four nucleotides on either side of the fragment start position, and in other just one nucleotide before and seven inside the fragment. The choice is dependent on the locations where it was found the nucleotide had most influence on fragmentation.

These figures indicate a wide range of different patterns between experiments and different labs. It is frequently the case that technical replicates which appear to have been processed at the same time have very similar characteristics. There are also specific patterns that occur in a number of different experiments. For example, there a number of experiments from the Myers/HudsonAlpha lab in Figure B-2 which show a similar characteristic, which is also seen in a set of *C. elegans* data from a completely different lab in Figure B-7c).

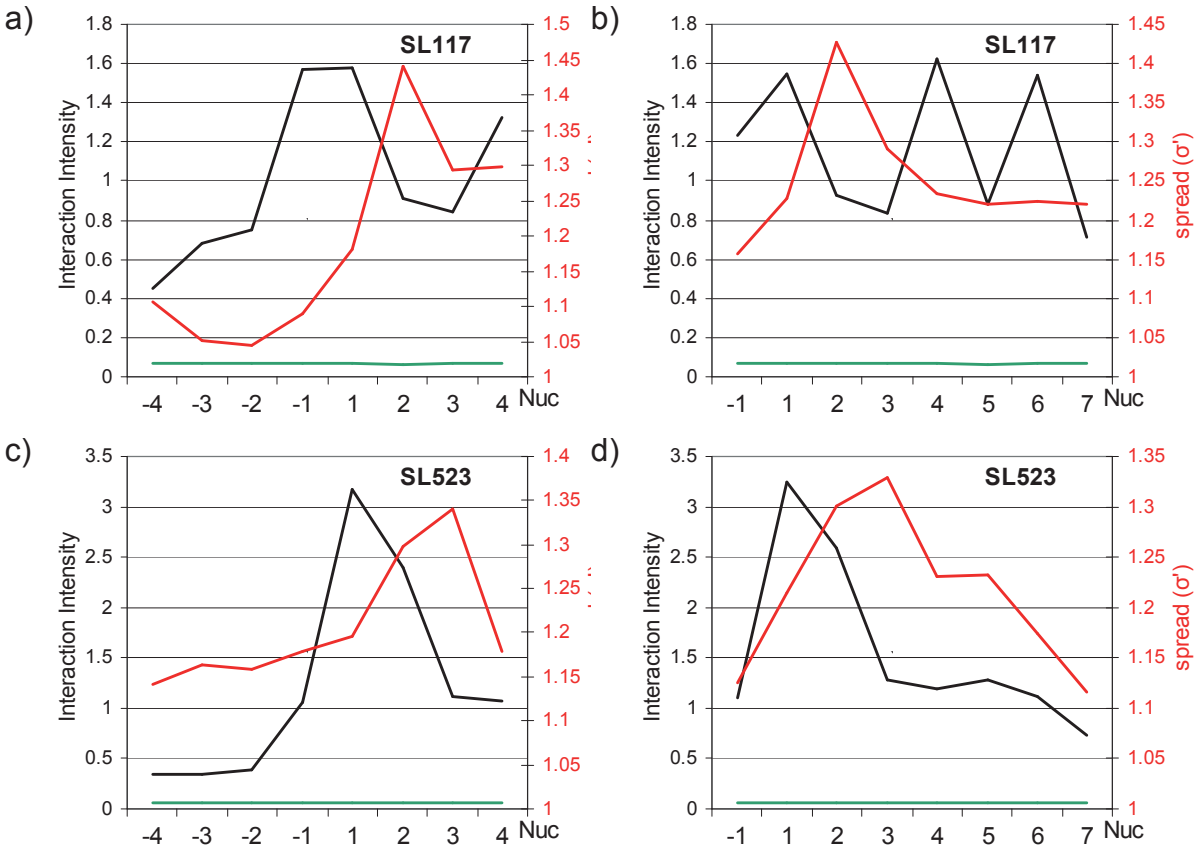


Figure B-1 Two technical replicates SL117 and SL523 show very different characteristics. These are the two datasets used in the main body of the thesis. a) and c) show the contributions for the eight nucleotides centred around the fragment start site. b) and d) are for the eight nucleotides starting from one nucleotide before the fragment start site so show the bias for the first seven nucleotides of the fragment. Black lines give a measure of the contribution to the likelihood of fragmentation using the average of the mutual information between the nucleotide and the probability of fragmentation for each of the other sets of nucleotides at the other positions. The red lines give a measure of the spread of the values. The green lines indicate the interaction intensity that would be expected if the fragment starts were uniformly distributed in the genome. SL117 shows a peak in the average around the fragment start position and also four and six nucleotides into the fragment. SL523 shows a peak at the first nucleotide of the fragment. In the pairs of graphs from each dataset there is a four nucleotide overlap, and the results in the overlapping regions broadly align, which provides a degree of validation for this technique.

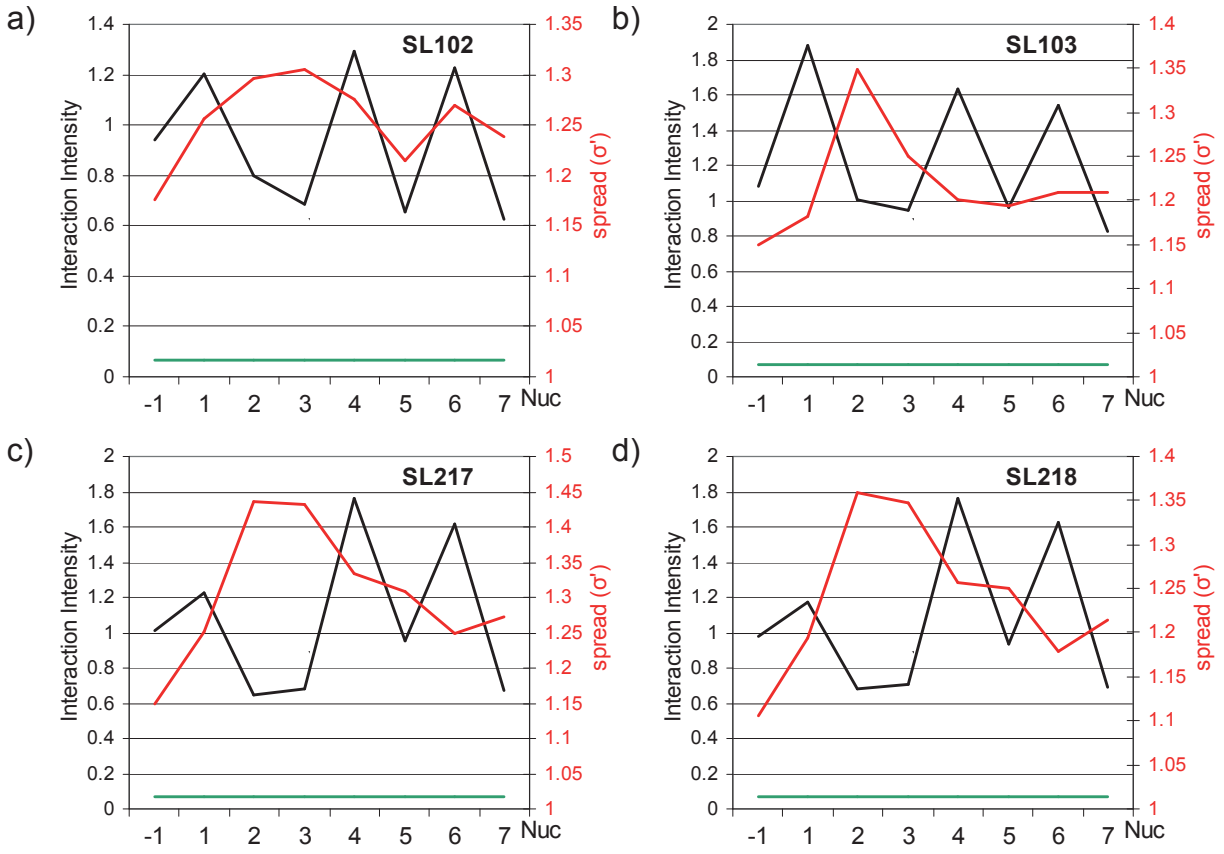


Figure B-2 Four early datasets from the Myers/HudsonAlpha lab show similar characteristics. These all show distinctive peaks in the interaction intensity at nucleotide positions one, four and six from the start of the fragment, which matches the results from the SL117 data.

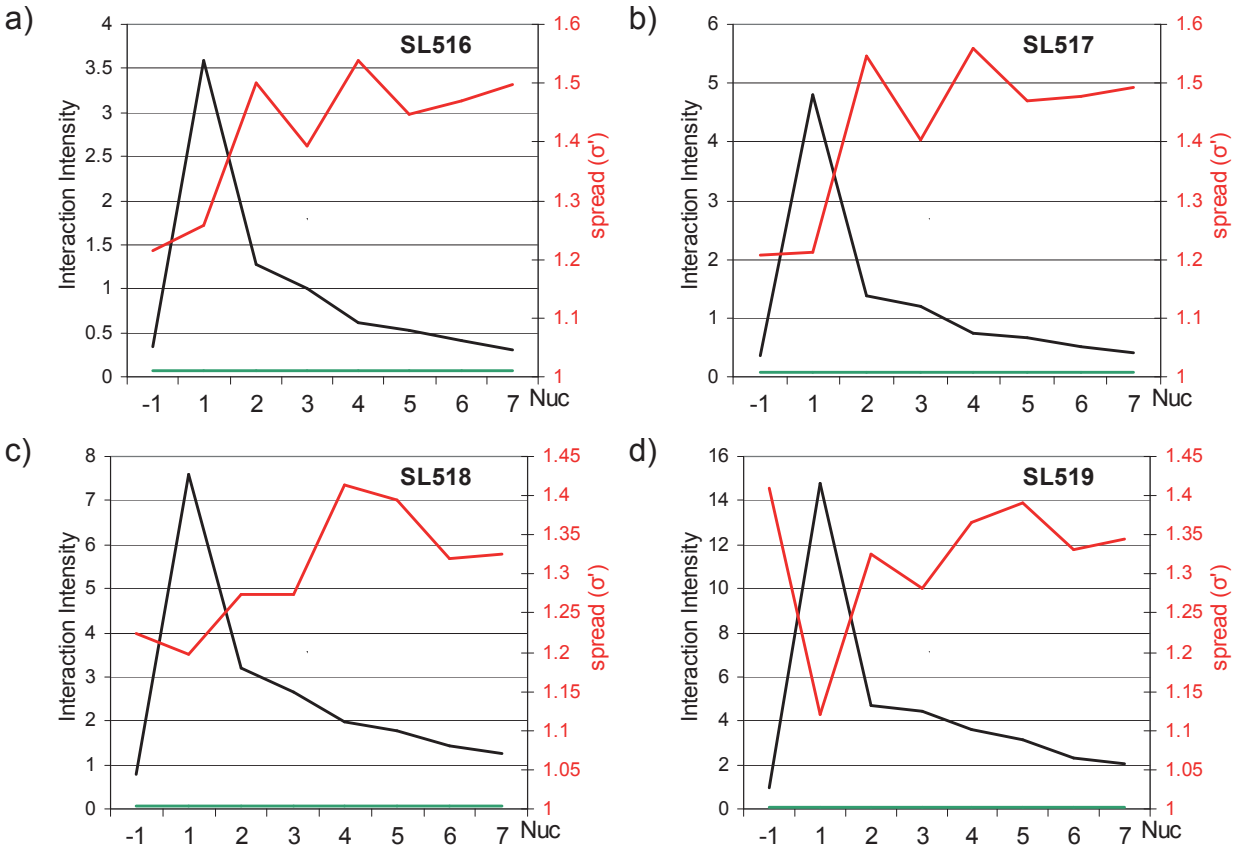


Figure B-3 Later Myers/HudsonAlpha results differ significantly from early results. These show a peak in the interaction intensity at the first nucleotide of the fragment and subsequently dropping away. The intensities are all significantly larger than the earlier set, suggesting that the nucleotides have a greater influence over the probability of DNA fragmentation. The spreads are similar to that seen in the previous graphs. These results are very similar to the SL523 data in Figure B-1.

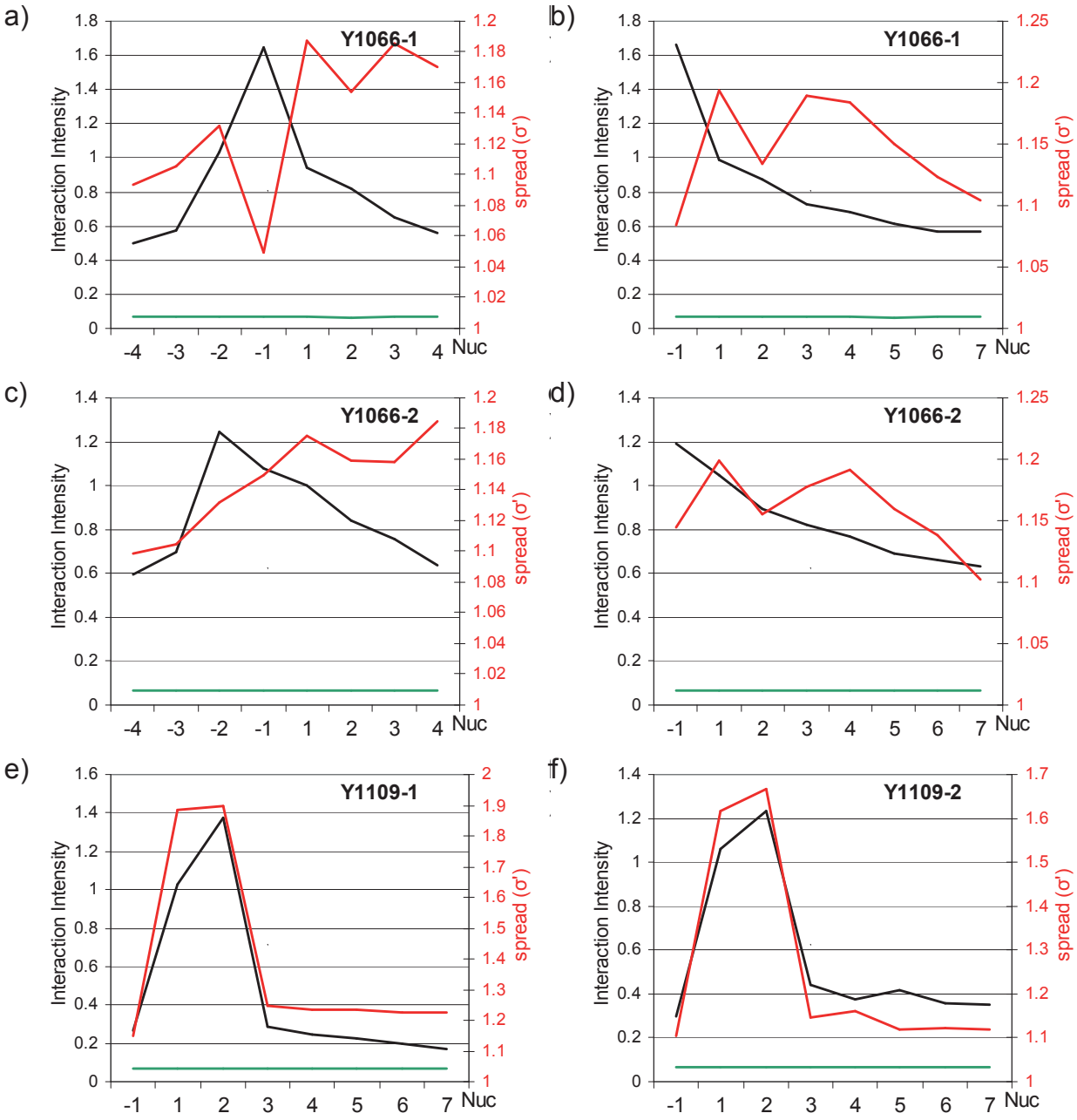


Figure B-4 Data from the Snyder/Yale lab show a variety of different characteristics. a) and c) are centred around the fragment start location and c) and d) are the same datasets shifted by three nucleotides to show that there is significant contributions to the sequence bias up to seven nucleotides in from the start of the fragment. e) and f) are two further datasets and cover the region starting from one nucleotide before the fragment start.

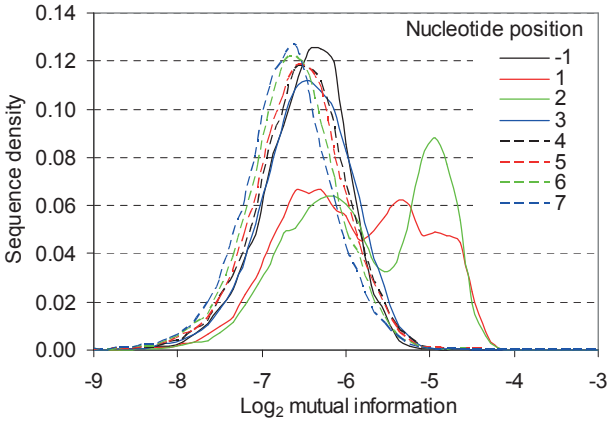


Figure B-5 Log mutual information distribution for Y1109-1 shows a complex picture underlies the simple interaction intensity and spread values The data shows the distribution of the log mutual information for the eight nucleotides that generated the summary in Figure B-4f)

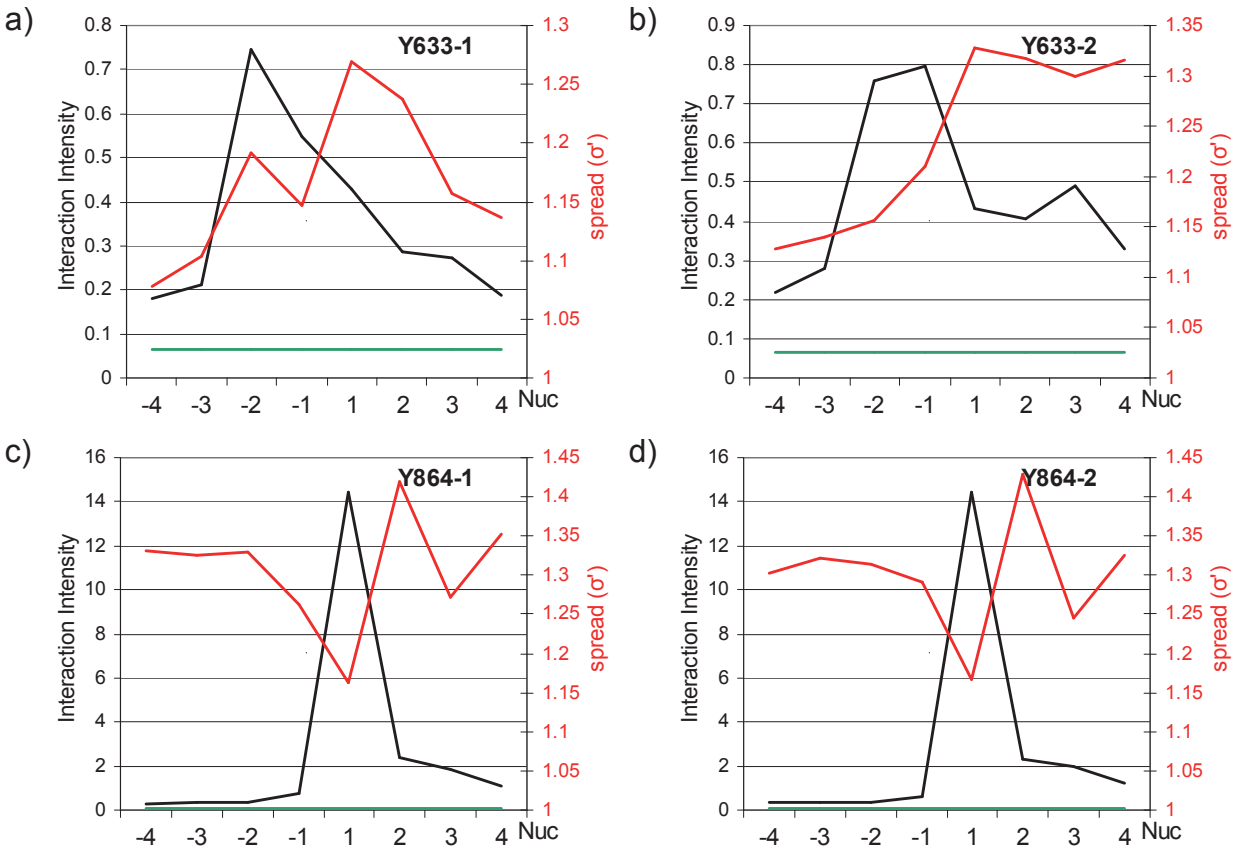


Figure B-6 A second example showing similarity between coincident technical replicates and differences between replicates from different dates. All four graphs are from input datasets using the K562 cell line from the Snyder/Yale lab. a) and b) were done at the same time and show similar characteristics to each other. c) and d) were also done together but at a later time, and show similar characteristics to each other, which are very different from the first technical replicates.

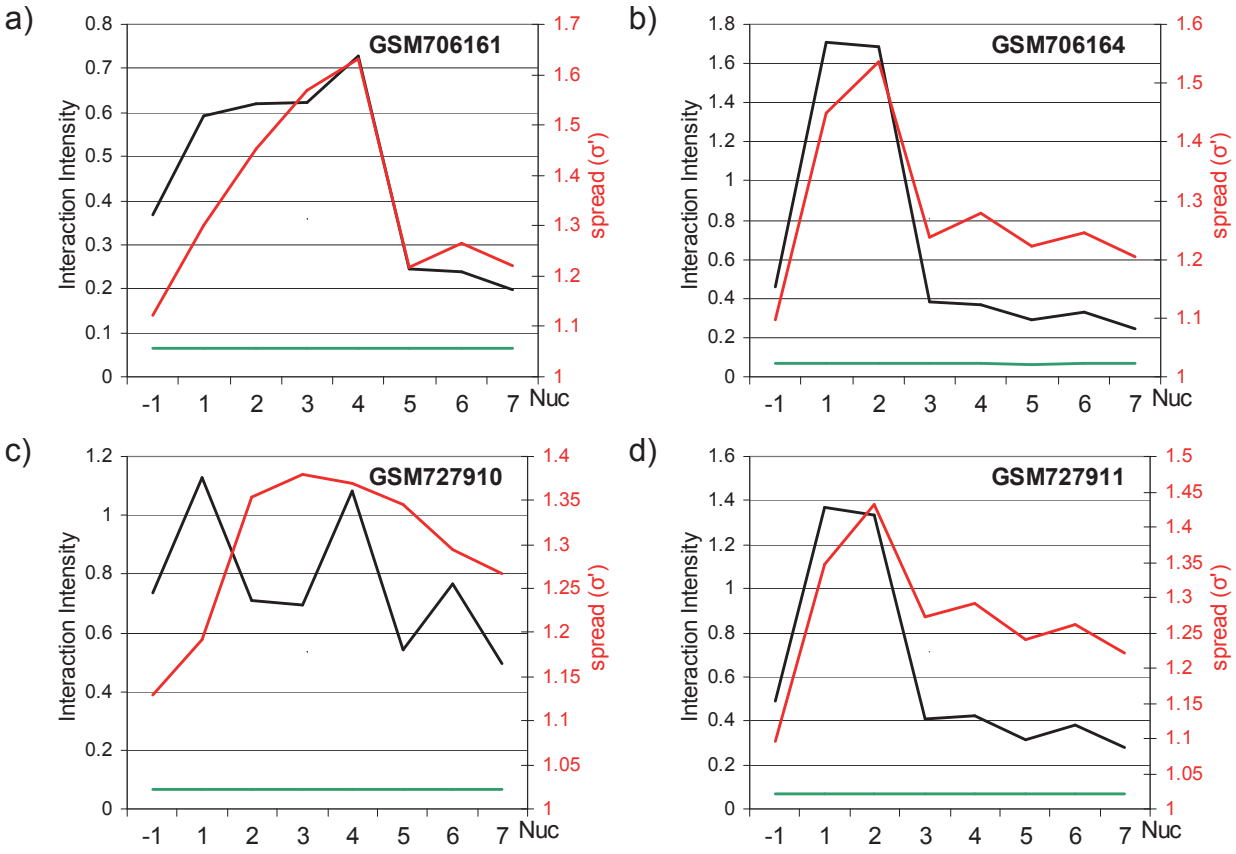


Figure B-7 Input fragments from *C. elegans* ChIP-seq experiments. These are four separate sets of input data which show a range of different characteristics. c), with peaks at nucleotide positions one four and six, is very similar to the characteristics seen in Figure B-2.

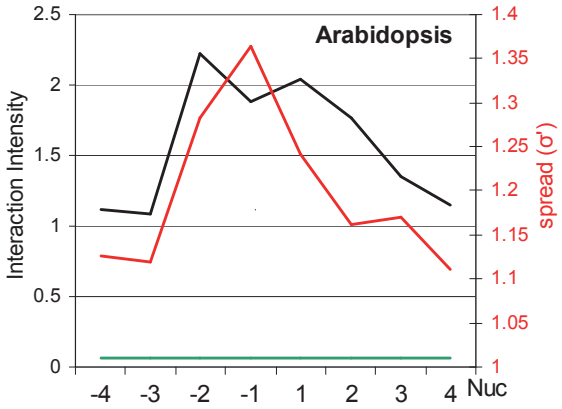


Figure B-8 Input fragments from an *Arabidopsis* ChIP-seq experiments.

Appendix C

Additional ChIP-seq model-fitting results

Sections 2.2.5 and 2.2.7 introduced the method for representing the sequence bias at the start of of DNA fragments from ChIP-seq experiments using one or more PCMs and section 2.3.5 and following examined a few examples of data obtained using this technique. This appendix provides further examples to support the analysis in the main body of the thesis.

C-1 Early Myers/HudsonAlpha lab results ‡

These datasets all show a very similar GC-rich pattern starting one nucleotide before the start of the fragment, similar to that seen for SL117 (Figure 2-11).

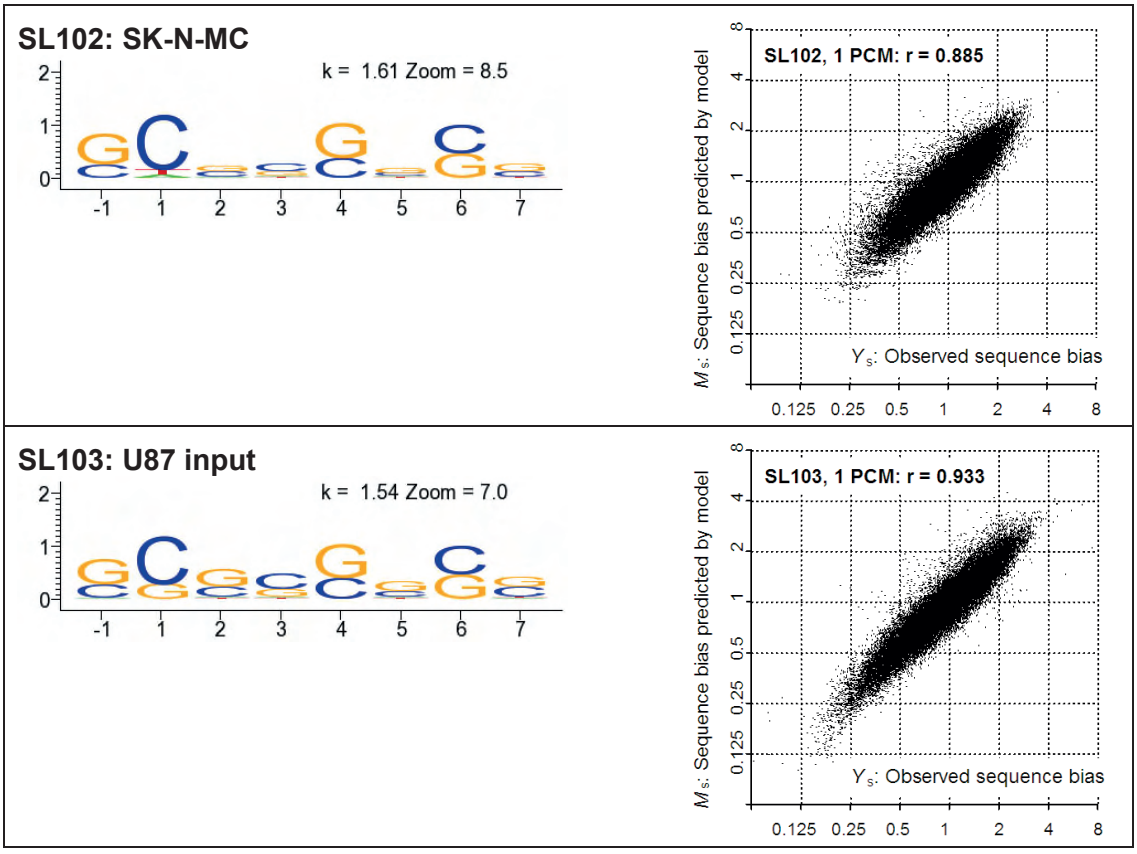


Figure C-1 Two early datasets from the Myers/HudsonAlpha lab.

C-2 A second pair of technical replicates with contrasting characteristics ‡

This is a second example which is similar to the SL117/SL523 pair (Figure 2-11) where the datasets are early (SL217/SL218) and later (SL516/517) input datasets using the same cell line, and show very different bias patterns.

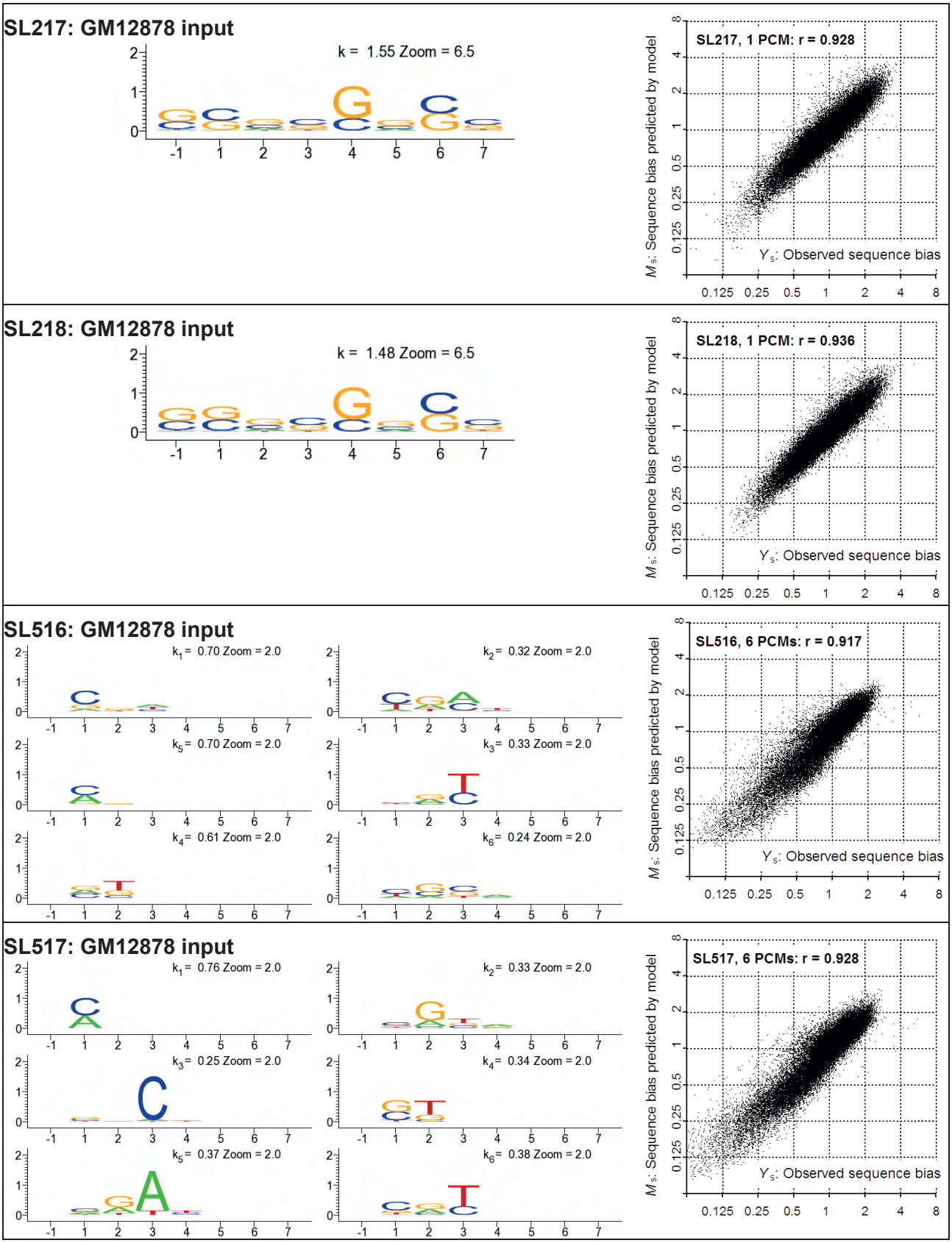


Figure C-2 Four input technical replicates from the same cell line with a significant range in sequence bias as indicated by the variety of PCMs

C-3 Late Myers/HudsonAlpha lab results ‡

These datasets all required a set of PCMs for optimal model fitting, as was the case for SL523. There is considerable variation in the PCMs between datasets.

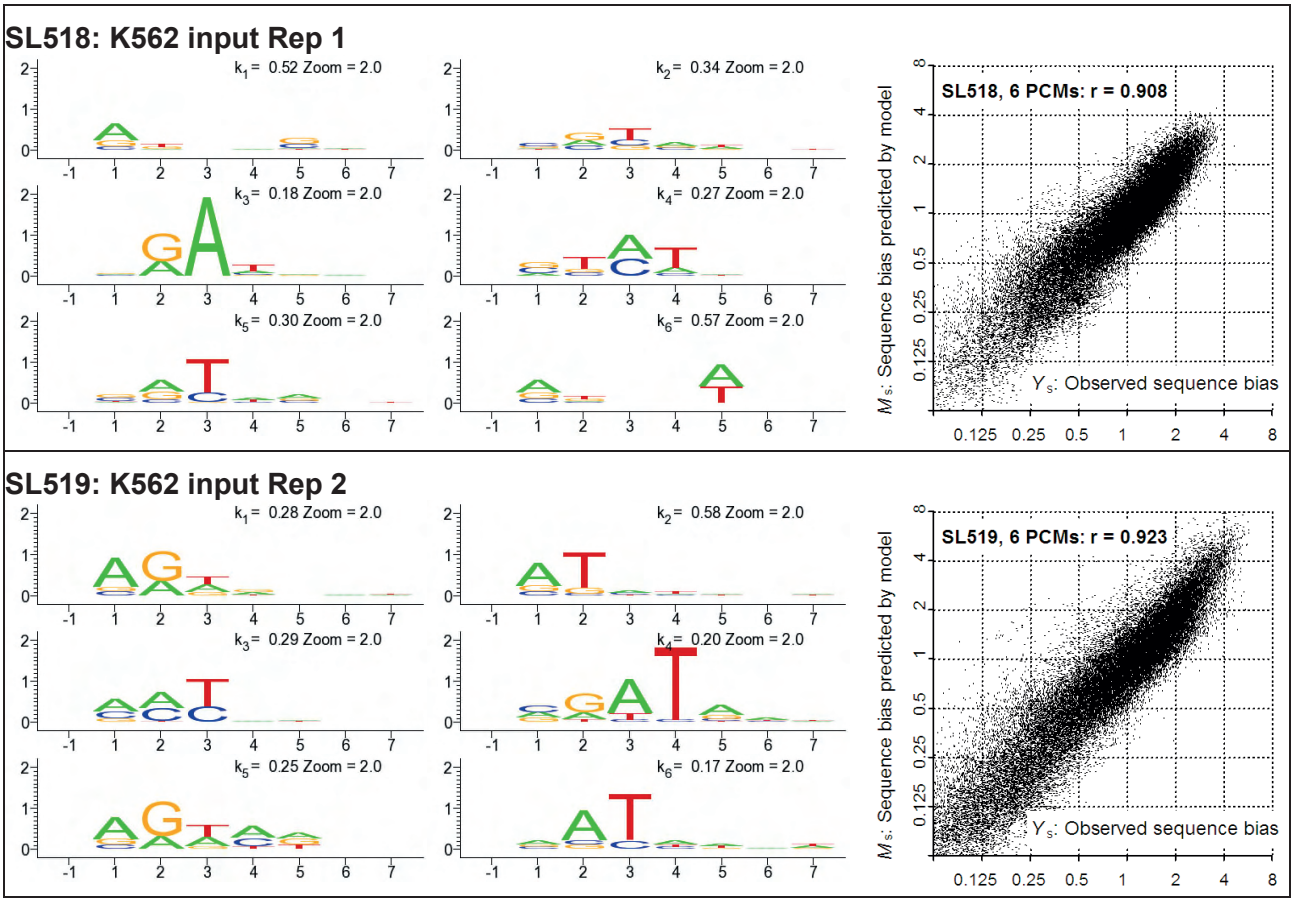


Figure C-3 Two later datasets from the Myers/HudsonAlpha lab showing a variety of multi-PCM characteristics.

C-4 Yale/UC-Davis/Harvard lab ChIP-seq data ‡

This section shows the results of carrying out a model fitting exercise on a selection of datasets from ChIP-seq experiments performed as part of the ENCODE project by the Yale/UC-Davis/Harvard lab and deposited in the Encode repository.

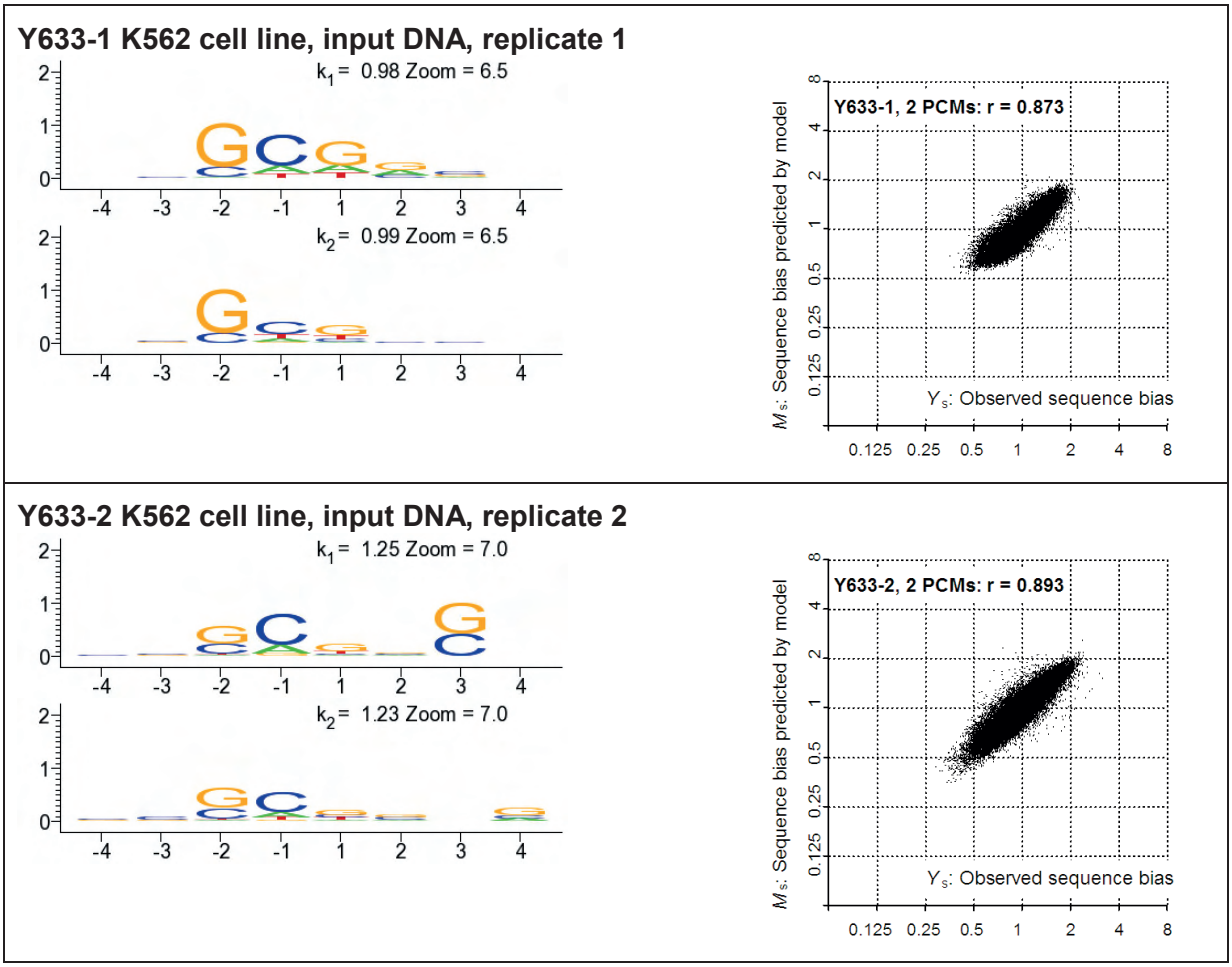


Figure C-4 Two sets of experimental data from the Yale/UC-Davis/Harvard lab which show CG bias around the fragment start side.

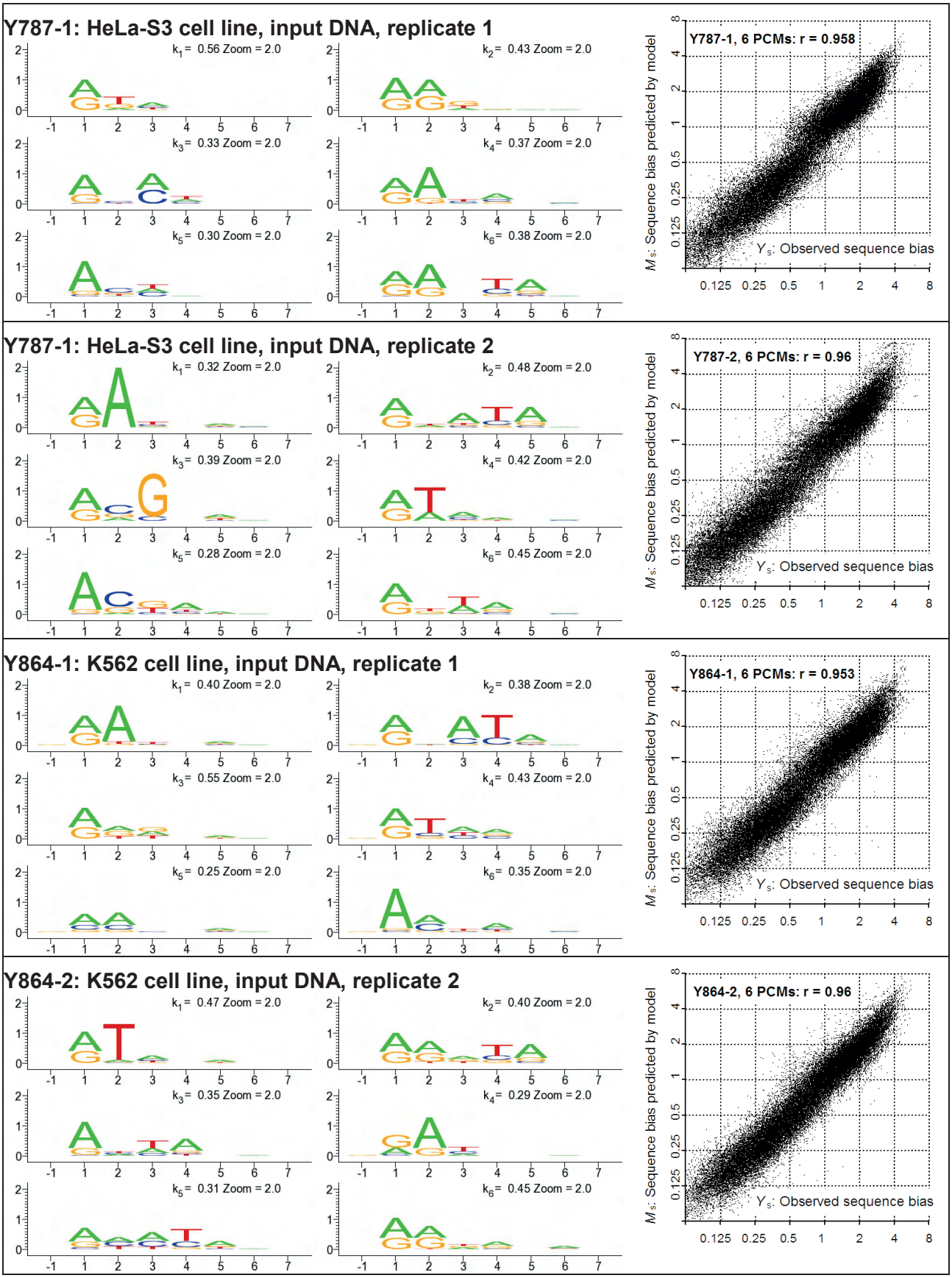


Figure C-5 Four sets of Yale/UC-Davis/Harvard data which show strong A/T bias only in the nucleotides within the fragment.

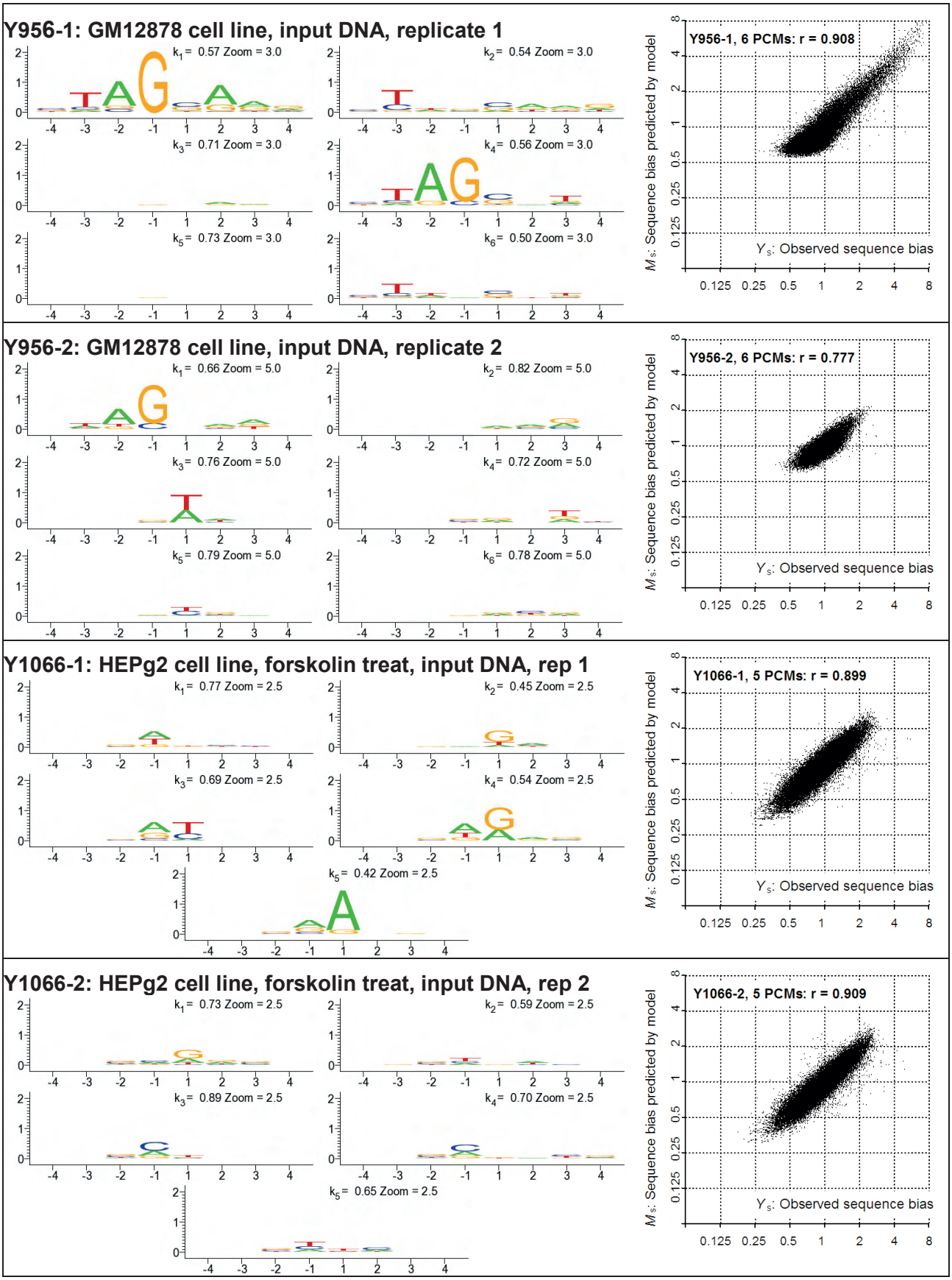


Figure C-6 Four Yale results with a range of different characteristics.

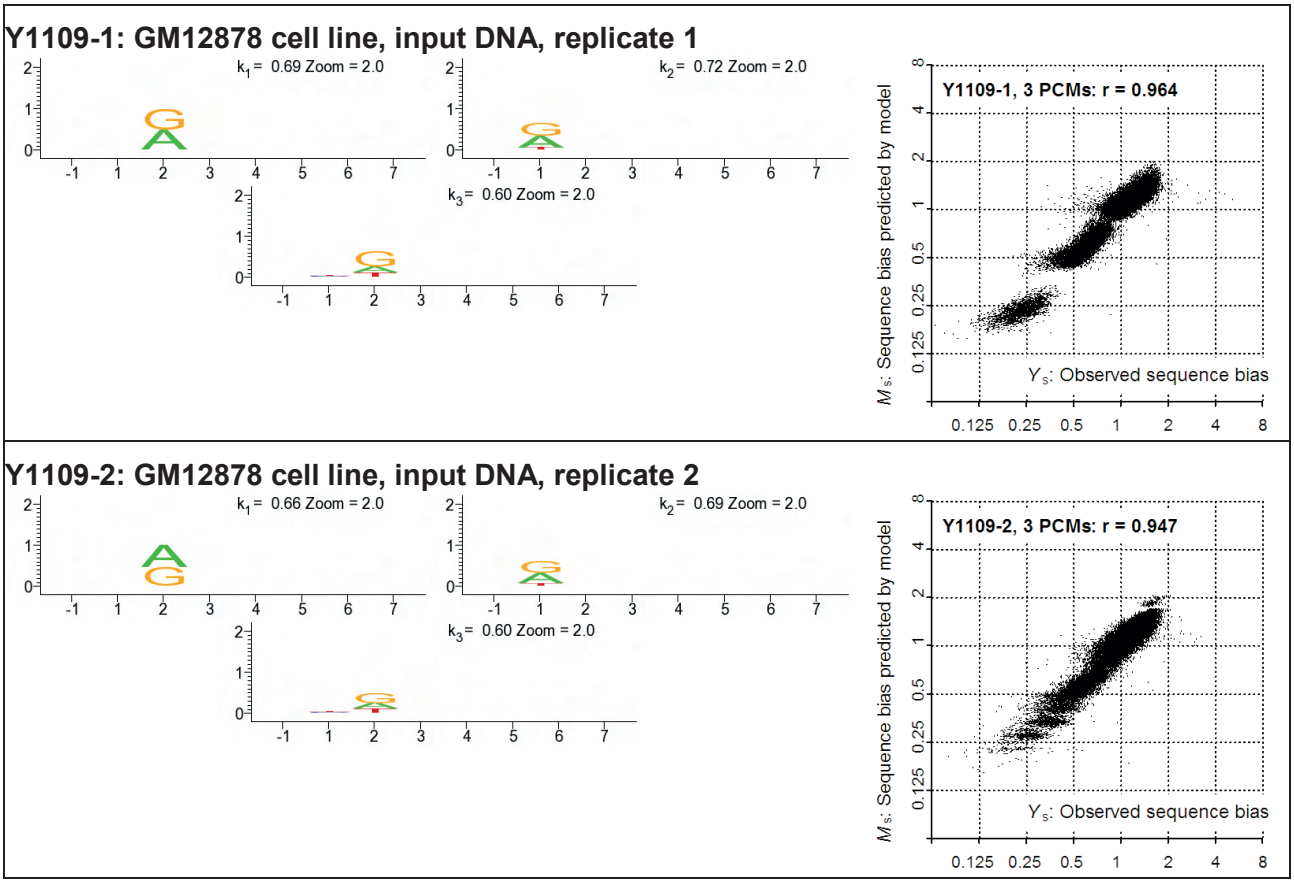


Figure C-7 Two Yale results with a distinctive AG bias in the first two nucleotides of the fragment

C-5 Cheung et al ‡ [20]

This is data that was used as part of an investigation into sequence bias in high throughput sequencing. The DNA was extracted from *C. elegans*, and the smaller genome meant that the fragment density is greater and there are fewer instances of each sequence in the genome. A threshold of 1000 instances of the 8-mers in the genome was used.

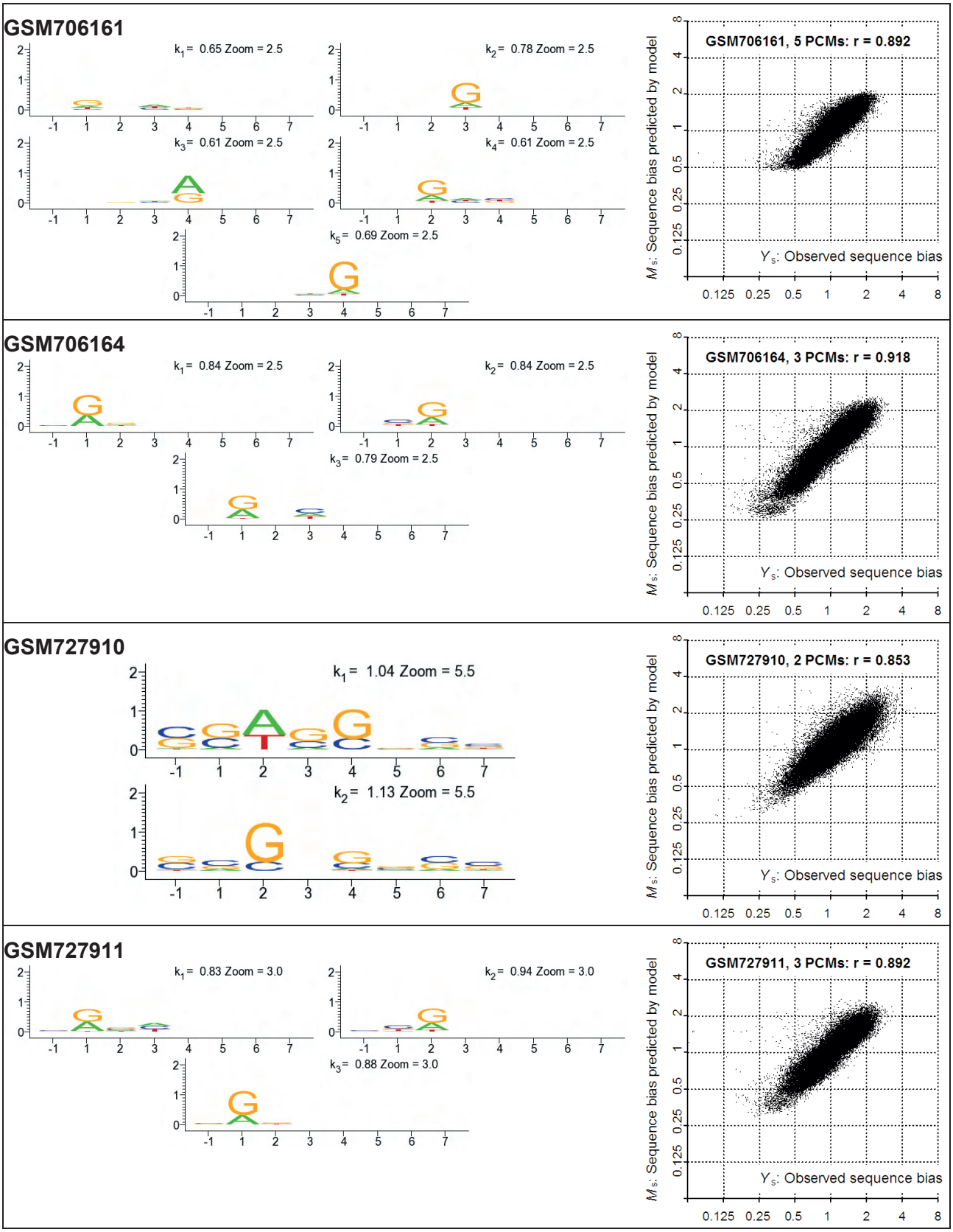


Figure C-8 A range of different nucleotide bias characteristics seen in *C. elegans* input data
Threshold T set to 1000 because the smaller *C. elegans* genome results in fewer instances of each sequence compared to *H. sapiens*

C-6 Arabidopsis dataset

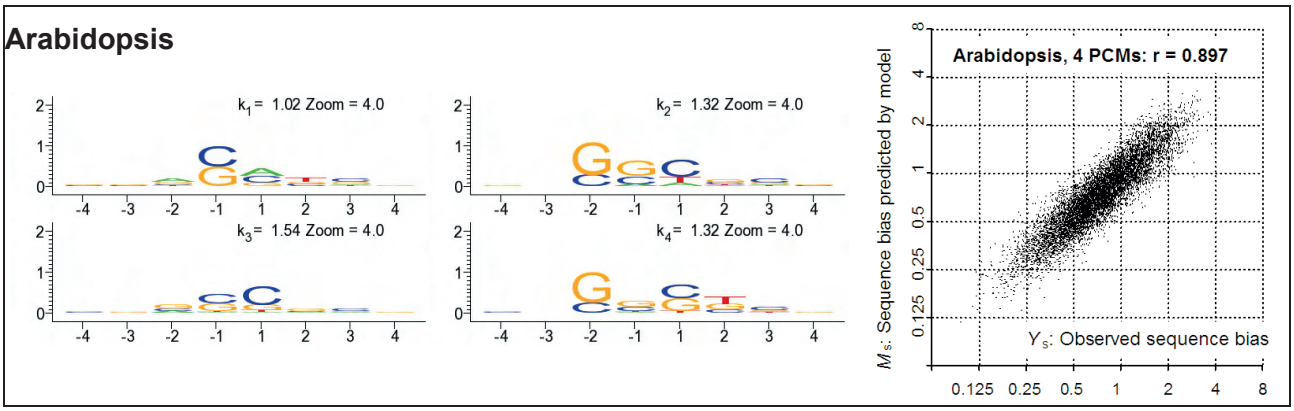


Figure C-9 Nucleotide bias characteristics from one sample of *Arabidopsis* input data
Threshold T set to 5000 because the smaller *Arabidopsis* genome results in fewer instances of each sequence compared to *H. sapiens*

Appendix D

Additional RNA-seq model-fitting results ‡

The following shows the result of fitting the model to a number of sets of RNA-seq data. These provide additional examples which show similar, but not identical characteristics to the results shown in Figure 3-3. In each case the results are shown when four PCMs are used to fit the bias for the first six nucleotides and then a single PCM for nucleotides seven onwards. The PCMs for the 3' end are shown as their reverse complement to highlight the similarity between the 5' and 3' biases when viewed in this way.

The model fitting and plotting of results uses the eight genes where there is the greatest RNA fragment density. In each case there is a graph showing the correlation between the measured number of fragment starts at each position within these genes (equivalent to the black line in Figure 4a and b) and the counts generated by the model (the red line in Figure 4a and b). In each case the data is matched to the full set of gene sequences rather than the representative set to cater for the possibility that the a significant number of sequences may have aligned to transcription variant that is not in the representative set.

The order of the PCMs that are generated by the model-fitting process is somewhat indeterminate. The PCMs have been reordered to highlight the similarity that exists between all of the sets of PCMs.

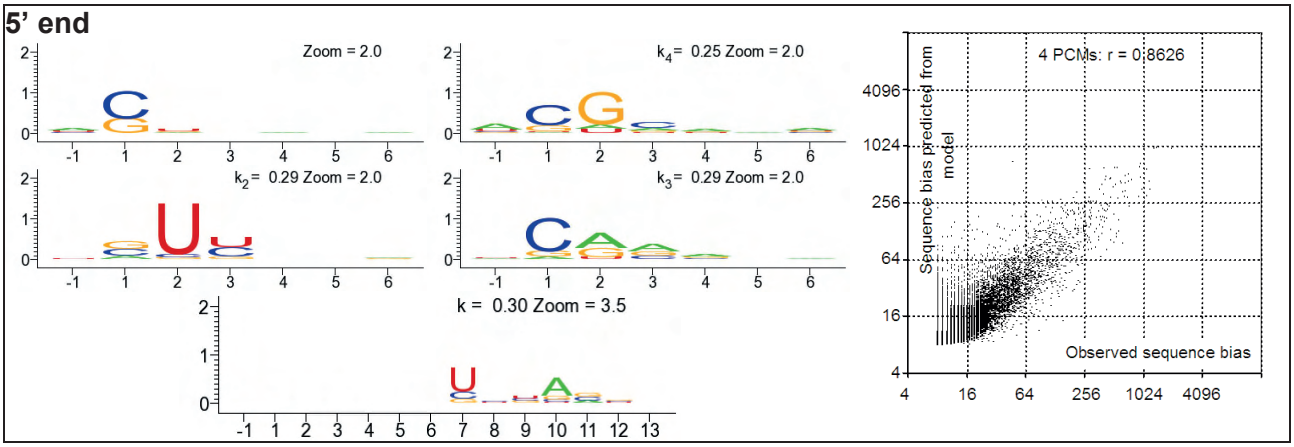


Figure D-1 RNA-seq model-fitting: GSM484895 5' end (*Homo sapiens*)

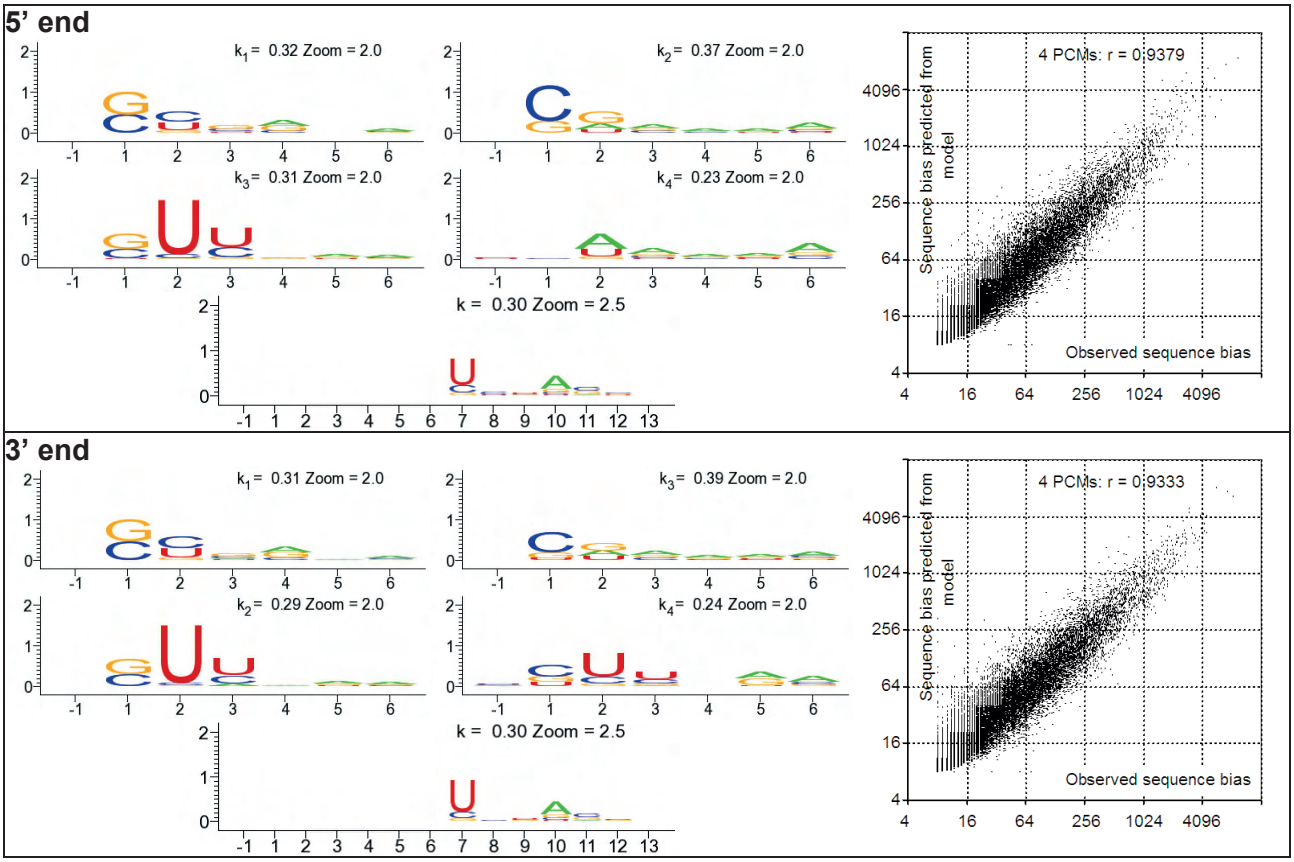


Figure D-2 RNA-seq model-fitting: Mouse skeletal data - Wold lab (SRX000352)

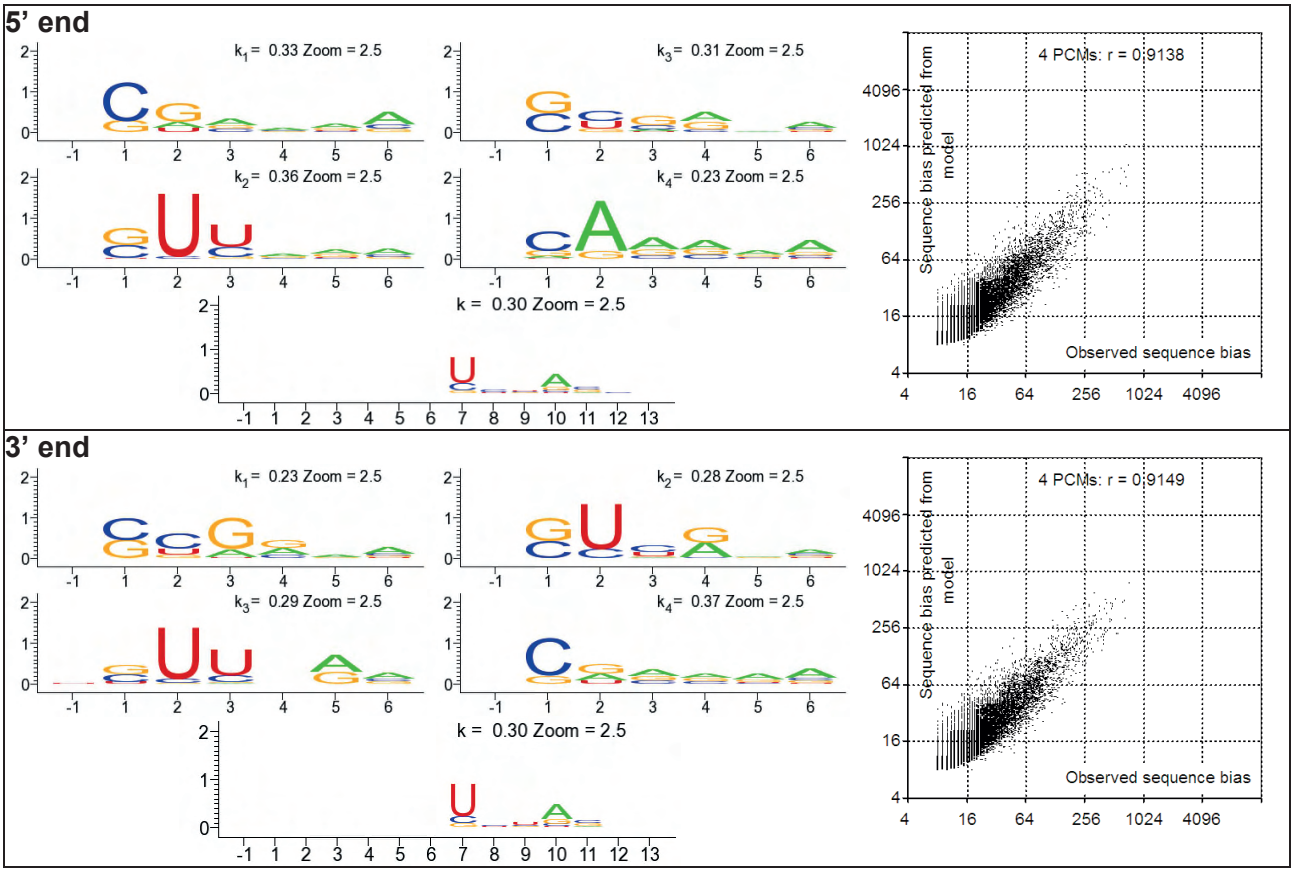


Figure D-3 RNA-seq model-fitting: Mouse brain- Wold lab (SRX001866)

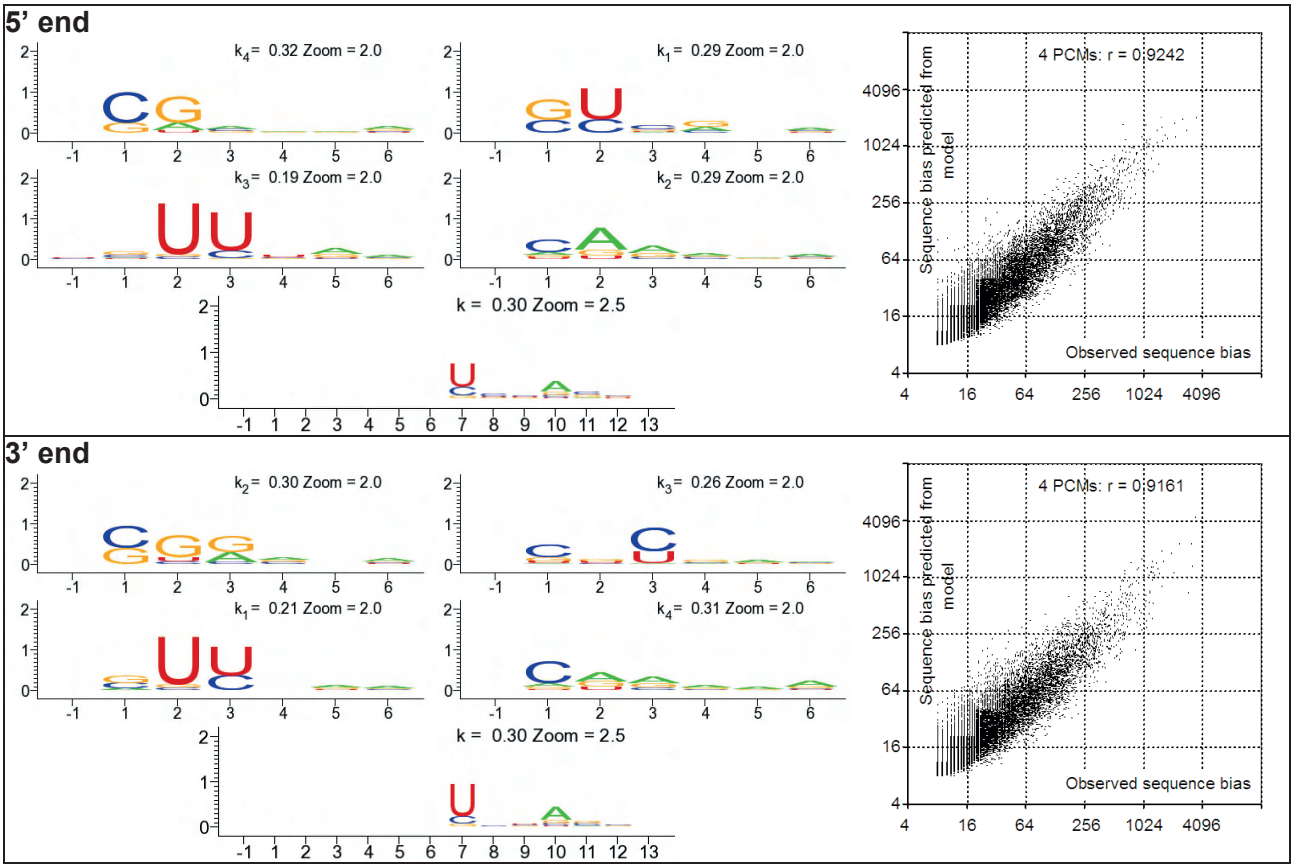


Figure D-4 RNA-seq model-fitting: Arabidopsis 24 hr: Replicate 1

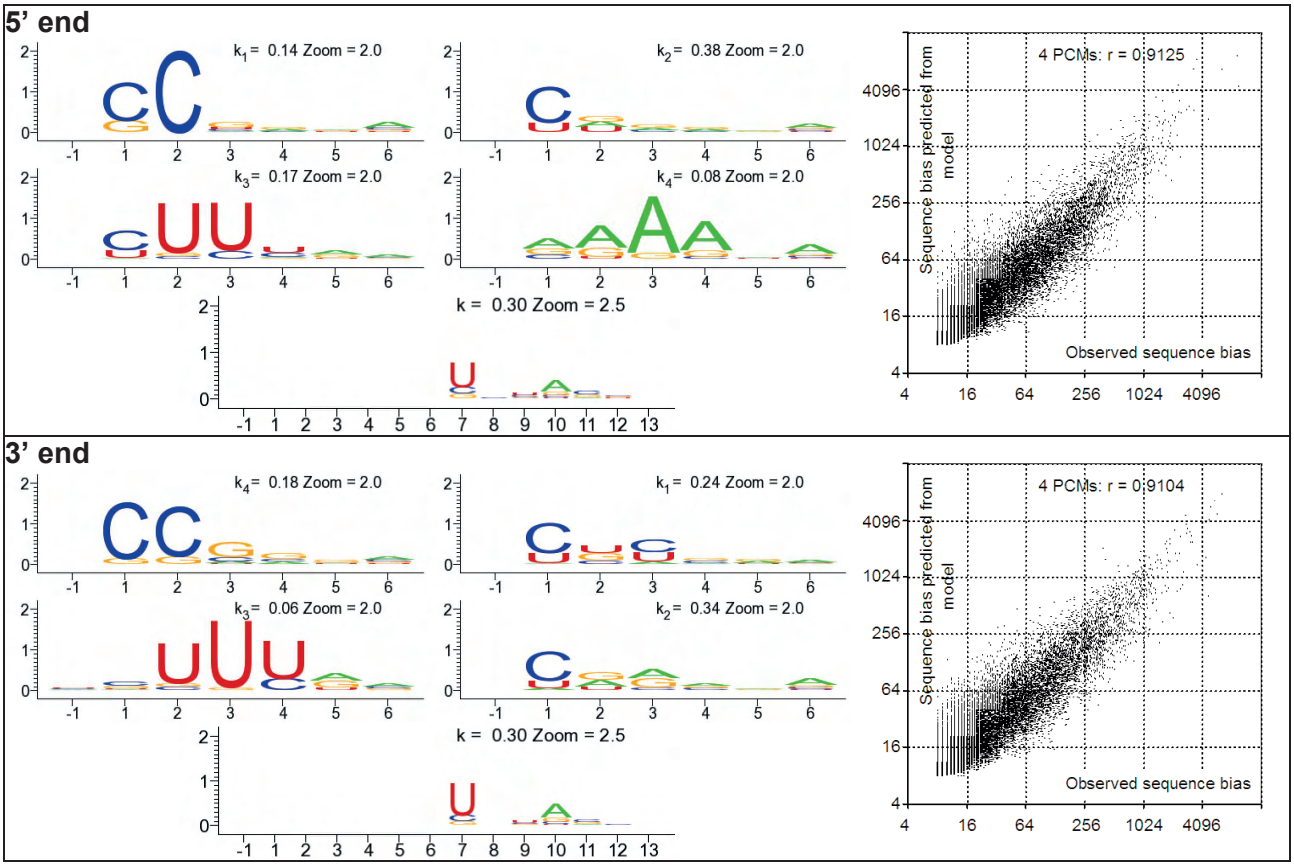


Figure D-5 RNA-seq model-fitting: Arabidopsis 24 hr: Replicate 2

Appendix E

Software architecture

A common software architecture was adopted for all of the analyses described in this document.

E-1 General principles

The general design objectives that were met by this architecture were:

1. **The core software should be efficient.** It needs to be able to process large amount of genomic data quickly and efficiently. This was achieved by writing all of the core software in C++.
2. **The core software should be multi-platform.** This was to allow the development and debugging of code locally on a PC and then the submission of jobs using the same code on the various high performance computing platforms available at the University of Warwick. The core C++ only used common libraries such as the Standard Template Library [92] to ensure the code was portable. The multi-platform Boost build software was used to build the software as it can run on all the target platforms using a common build definition file.
3. **It should be possible to run the code using Graphical User Interfaces (GUI)**. The nature of the investigation meant that it was frequently necessary to return to using the software tools after an extended period of time. This is made easier by the use of GUIs to set up the parameters, removing the need to remember complex command line formats. The cisGenome software incorporates GUIs for setting up the parameters for command line tools. These GUIs were extended for additional code that was written which was an extension of the cisGenome functionality. Microsoft Excel was used to create simple GUIs for running other software. In both cases the GUIs were designed to allow easy access to the command lines they had generated so that they could be used as templates for running the same code on other platforms.
4. **Code releases should be managed.** This was done through the use of the open source SVN versioning software which allows the changes to the source code to be recorded over time. This makes it possible to reproduce if necessary the

results that were obtained using early versions of software and investigate unexpected changes that might appear in the way that the software runs. This also provides a mechanism for transferring source code between platforms in a controlled way.

5. **It should be easy to visualise the results graphically.** It is frequently the case that patterns in genomic data become clearer when they are able to be visualised, which is then the first stage of a more analytic analysis of the data. The ability of the cisGenome software suite to display distributed genomic data was one of the reasons why it was chosen for the project. Excel spreadsheets were also used to display results in both tabular and graphical format.
6. **There should be simple data transfer processes.** The need to transfer data between different applications and different processes, and the need for flexibility in order to be able to test out alternative algorithms led to the use of comma separated variable format text files for the storage and transfer of both bulk data such as genomics data and also configuration and model fitting parameters. The also enables simple transfer of intermediate results between different processing platforms.
7. **There should be a good software development environment.** The Visual C++ development environment was used for the development and debugging of the C++ code, prior to its use on both Windows and also Apple and Linux platforms. The Boost build software was used to build the software on the Windows platform prior to it being transferred and built on other platforms. All of the GUI tools create the command lines for their associated tools in such a way that the command lines could be copied and pasted into the development environment for testing or into scripts to be run on other platforms.

E-2 Software architecture

Figure E-1 gives an overview of the program modules used for the analysis described in this document. The usual pattern was that a problem was initially investigated in the Microsoft Windows environment and then when the process had been established, the analysis would be performed using the complete set of observed data files on the high performance server platforms that are available at the University of Warwick.

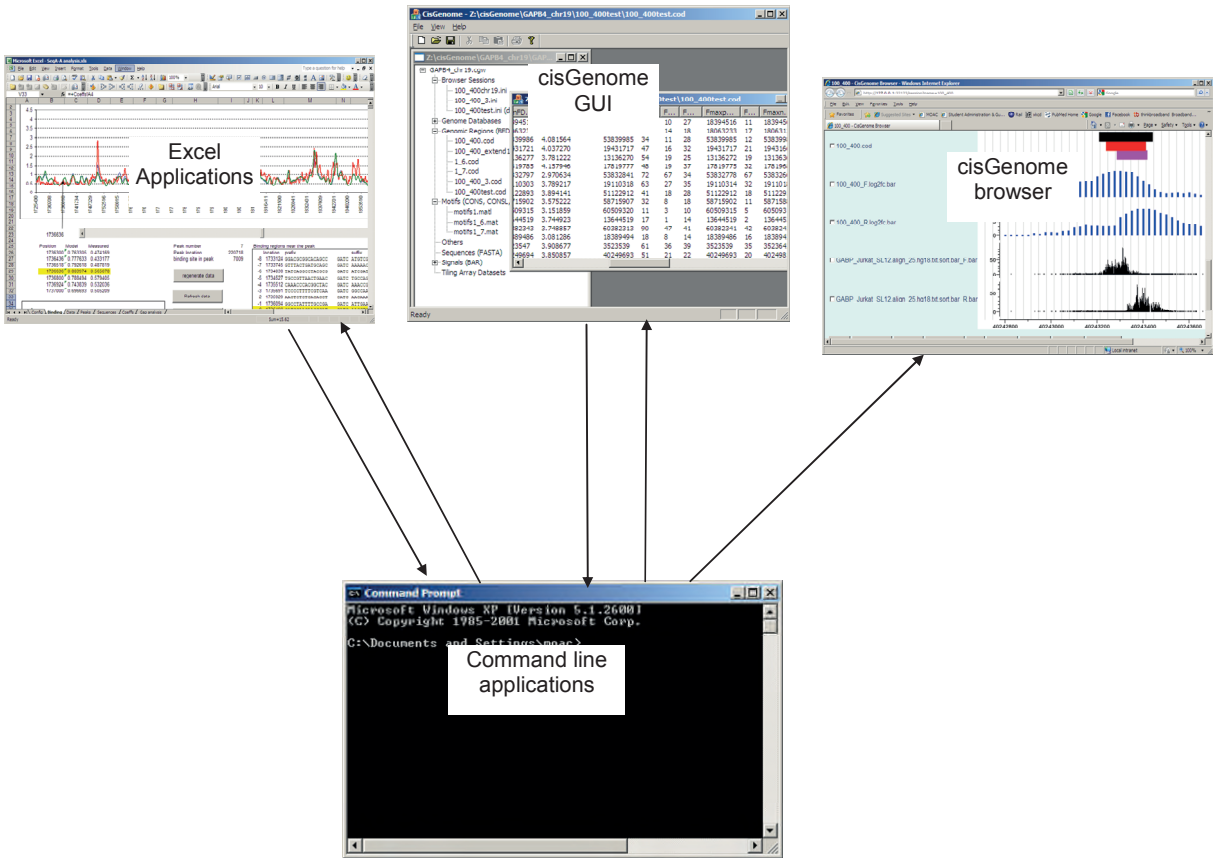


Figure E-1 Interactions between program modules. Both Excel applications and the cisGenome GUI are able to drive Windows command line applications and load the results from these applications for display. They can also generate the command lines for running the programs on other platforms. The cisGenome browser is able to load and display PCM files and genome sequence related data.

Appendix F

Ancillary algorithms

The following sections provide additional background to some of the algorithms that were used at various times within this research.

F-1 Assessing significance using Pearson's coefficient and the Fisher transformation

A frequent problem that arises when investigating correlation between two datasets is assessing the significance of the result. One way of making this assessment is to consider the likelihood that the correlation that is seen could have arisen given the null hypothesis that the two distributions are unrelated and uncorrelated. This problem was addressed by Fisher [33] who proposed a transform that could be applied to the Pearson product-moment coefficient which provides a measure of the correlation between two sets of data

The Pearson coefficient for two datasets can be calculated as follows:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})}} \tag{8.1}$$

x_i and y_i are the paired elements from the two datasets, each of length n, and $\bar{x}$ and $\bar{y}$ are the means of the two datasets. The definition of the Fisher transformation $F(r)$, and the standard error for this distribution is then given by:

$$F(r) = \frac{1}{2}\ln\frac{1+r}{1-r} \qquad SE = \frac{1}{\sqrt{n-3}} \tag{8.2}$$

This can be converted to a z-score which tends to a normal distribution with a mean of zero and a standard deviation of one using:

$$z = \frac{F(r) - F(\rho_0)}{1/\sqrt{n-3}} \tag{8.3}$$

The cumulative normal function is then used to calculate a p value associated with a value in this distribution, giving the probability of finding a value as least as big as the value, either on one side (one tailed) or on both sides (two tailed) of the normal distribution.

F-2 Calculation of cumulative normal values for large z

A problem arose of wanting to calculate the p-value for a z-value that was more than 200 standard deviations from the mean of a normal distribution, such that the limited

precision of the calculations in software such as Excel rendered the result as zero, rather than an extremely small non-zero value. In some ways this is somewhat academic because it corresponds to an extremely unlikely event. Nevertheless, the following was derived in order to calculate the value to the nearest order of magnitude, rather than be limited to stating that the value was essentially zero.

The normal distribution $\phi(x)$ with a mean of zero and standard deviation of one is defined by the following equation:

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \tag{8.4}$$

The $1/\sqrt{2\pi}$ factor is a normalisation term to ensure that the total area under the curve equals one. The probability of getting a value greater than x is then

$$\begin{aligned} \Phi(x) &= \int_x^\infty \phi(x)dx \\ &= \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-x^2/2} dx \\ &\approx \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \left(\frac{1}{x} - \frac{1}{x^3} + \ldots \right). \end{aligned} \tag{8.5}$$

The first term in the series is the only significant term for large values of x, so this can be approximated as

$$\begin{aligned} \Phi(x) \approx \frac{e^{-x^2/2}}{x\sqrt{2\pi}} . &= \frac{10^{-x^2/2\ln(10)}}{x\sqrt{2\pi}} \\ &= 10^{-x^2/2\ln(10)} * 10^{\log_{10}\left(1/x\sqrt{2\pi}\right)} \\ &= 10^{\left(-x^2/2\ln(10) - \log_{10}\left(x\sqrt{2\pi}\right)\right)} \\ &= 10^y \quad \text{where} \quad y \approx -0.21715x^2 - \log_{10}\left(2.507x\right) \end{aligned} \tag{8.6}$$

The value of y can be calculated for values of x that are significantly greater than 200 without incurring problems with arithmetic precision.

Appendix G

Co-authored journal publications

G-1 An alignment-free model for comparison of regulatory sequences [55]

This paper proposes a new method for aligning genomic sequences from different organisms in order to identify orthologous regulatory regions and arose out of the PhD studies of Hashem Koohy who completed his PhD at Warwick in 2010. The motivation for this method is that other alignment methods that have been developed for other applications, such as those used for the identification of transcription factor binding sites are not well suited to the identifying orthologous regulatory regions. This is because regulatory regions are more extensive, the sequence similarity is not necessarily as constrained, and there may be some reordering of the sequences within the regions.

My contribution to this paper arose out of the work I had carried out as part of my PhD to develop a computing infrastructure that allowed genetic sequences to be submitted to a central server which was able to search for matches between the sequences and the descriptions of transcription factor binding site sequences contained within the database.

These descriptions are in the form of Position Specific Scoring Matrixes (PSSMs) and the experience of working with Hashem with these PSSMs greatly informed my understanding of the strengths and weaknesses of using PSSMs which contributed to the way that these are used in the modelling of bias in ChIP-seq and RNA-seq data.

In addition, there was some reuse of some specific software components and general algorithmic techniques between the two developments.

G-2 CisGenome Browser: A flexible tool for genomic data visualization [47]

At the start of the research described in this thesis a review was undertaken of the various options for processing and visualising the genomics data such as ChIP-seq and RNA-seq data. The conclusion of this study was that the cisGenome software suite [46] provided many features that would support the analysis to be carried out. Some of its advantages include:

a) It provides a flexible framework with a powerful Graphical User Interface (GUI) for managing data files that are used and created when working with ChIP-seq and RNA-seq data.
b) It provides a flexible visualisation tool that allows the results to be viewed graphically and efficiently, including the ability to relate the results of an analysis to genomic features such as the location of genes
c) It provides an integrated multi-platform set of software tools that can perform many of the data analysis tasks required during this investigation.

The authors of the software kindly provided the source code which was then used as a foundation for significant changes and improvements that were needed as the PhD progressed. Changes introduced included:

a) The ability to import a wider range of data file formats
b) Improved smoothing and averaging of data during visualisation
c) Considerable GUI usability improvements
d) Software architecture changes to allow the visualisation component to be run on operating systems other than Microsoft Windows®

These changes were done in collaboration with the original software authors and they were sent copies of the new code after the changes that were made. Some of the changes and principles behind the changes were then incorporated into the next major release of the software for which this application note was published.

G-3 Dynamic distribution of SeqA protein across the chromosome of Escherichia coli K-12 [86]

The analysis of SeqA binding in *E. coli* described in Chapter 5 arose out of an invitation to examine some ChIP-chip data that had been produced to investigate the binding of the SeqA protein to the genome in *E. coli*. Some of the initial results of this investigation were included in this paper that was published in mBio.

The initial results published showed that there is a tendency for SeqA to be bound at locations that are separated by 9 or 10 nucleotides (Figure G-1). The model fitting techniques described in this document were subsequently developed to investigate this relationship in more detail

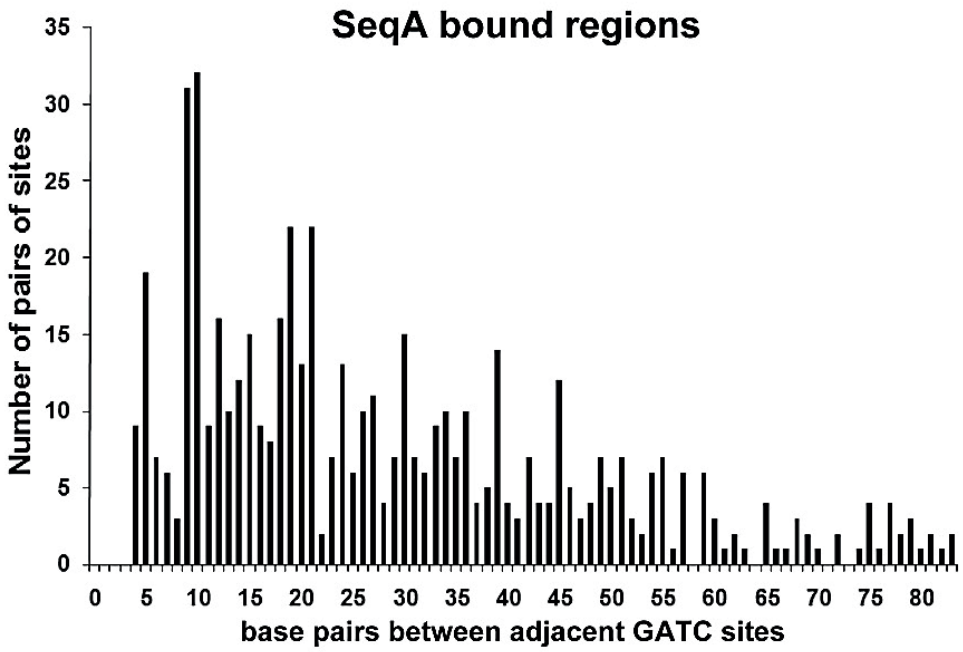


Figure G-1 Spacing of adjacent GATC motifs. This is a histogram of the spacing across the whole E. coli genome and in genomic regions bound by SeqA. The data show that, *in vivo*, a gap of close to 10 or 20 nucleotides between GATC motifs is most favourable for SeqA binding.

G-4 Variable structure motifs for transcription factor binding sites [82]

At various times during the period of this PhD there were a number of opportunities to work with other teams on the analysis of ChIP-seq data. One such opportunity was to work with *John Reid* and *Kenneth J Evans*, then at Birkbeck College, London on techniques for identifying transcription factor binding sites where there are multiple conserved regions within the binding sites.

This collaboration arose out of early investigations into transcription factor binding sites that was carried out as part of this research. At this stage studies were made of the detailed binding site data that is held in the TRANSFAC database [70] which are use to create the PSSMs that are published in this database. The results of my work were then incorporated into the paper (Figure G-2). The techniques used to generate the multiple sequence patterns represented by these PSSMs in this thesis are very different to those used in this paper.

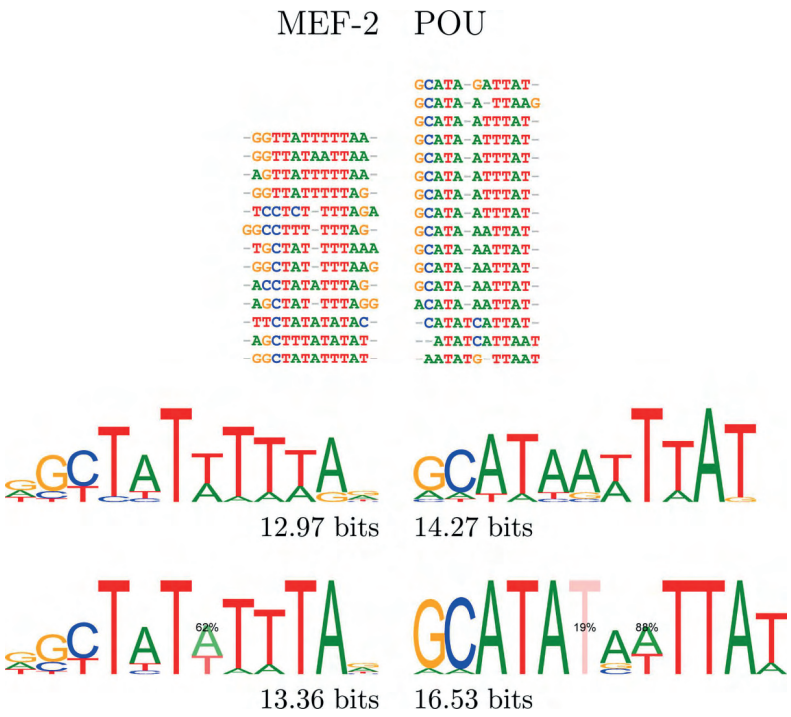


Figure G-2 Analysis of TRANSFAC binding sites . Realignment of sequences used within TRANSFAC to define PWMs which incorporate optional gaps. Left: An alignment, a standard PWM and a gapped PWM for the monomer transcription factor MEF-2. Additional gaps improve the alignment right across the motif, especially the well conserved TA motif that is not apparent in the ungapped alignment. Right: An alignment, a standard PWM and a gapped PWM for the homodimer transcription factor POU. Additional gaps improve the alignment of the conserved ATA and TTA motifs. The realignments show a significant proportion of sites both with and without gaps. The upper logos show the original TRANSFAC motifs and their information content in bits. The lower logos show the motifs after the addition of gaps, indicated by the percentage of sequences where a nucleotide is present.

Bibliography

1. **Finishing the euchromatic sequence of the human genome.** *Nature* 2004, **431:**931-945.
2. Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C, Gnirke A: **Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries.** *Genome biology* 2011, **12:**R18.
3. Auerbach RK, Euskirchen G, Rozowsky J, Lamarre-Vincent N, Moqtaderi Z, Lefrancois P, Struhl K, Gerstein M, Snyder M: **Mapping accessible chromatin regions using Sono-Seq.** *Proceedings of the National Academy of Sciences of the United States of America* 2009, **106:**14926-14931.
4. Badis G, Berger MF, Philippakis AA, Talukder S, Gehrke AR, Jaeger SA, Chan ET, Metzler G, Vedenko A, Chen X, et al: **Diversity and complexity in DNA recognition by transcription factors.** *Science (New York, NY)* 2009, **324:**1720-1723.
5. Bailey TL, Elkan C: **Unsupervised learning of multiple motifs in biopolymers using expectation maximization** *Machine Learning* 1995, **21:**51-80.
6. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Marshall KA, et al: **NCBI GEO: archive for high-throughput functional genomic data.** *Nucleic acids research* 2009, **37:**D885-D890.
7. Barski A, Zhao K: **Genomic location analysis by ChIP-Seq.** *Journal of cellular biochemistry* 2009, **107:**11-18.
8. Batchelor AH, Piper DE, de la Brousse FC, McKnight SL, Wolberger C: **The structure of GABPalpha/beta: an ETS domain- ankyrin repeat heterodimer bound to DNA.** *Science (New York, NY)* 1998, **279:**1037-1041.
9. Bauer T, Eils R, Konig R: **RIP: the regulatory interaction predictor--a machine learning-based approach for predicting target genes of transcription factors.** *Bioinformatics (Oxford, England)* 2011, **27:**2239-2247.
10. Beadle GW, Tatum EL: **Genetic Control of Biochemical Reactions in Neurospora.** *Proceedings of the National Academy of Sciences of the United States of America* 1941, **27:**499-506.
11. Berget SM, Moore C, Sharp PA: **Spliced segments at the 5' terminus of adenovirus 2 late mRNA.** *Proceedings of the National Academy of Sciences of the United States of America* 1977, **74:**3171-3175.

12. Berk AJ, Sharp PA: **Sizing and mapping of early adenovirus mRNAs by gel electrophoresis of S1 endonuclease-digested hybrids.** *Cell* 1977, **12:**721-732.
13. Binkley D, Dessy R: **Data manipulation and handling.** *Journal of Chemical Education* 1979, **56:**148.
14. Boveri T: **Zellenstudien II: Die Befruchtung und Teilung des Eies von Ascaris megalocephala.** *Jenaer Zeitschrift für Naturwissenschaft* 1888, **22:** 685–882.
15. Brendler T, Sawitzke J, Sergueev K, Austin S: **A case for sliding SeqA tracts at anchored replication forks during Escherichia coli chromosome replication and segregation.** *The EMBO journal* 2000, **19:**6249-6258.
16. Bullard JH, Purdom E, Hansen KD, Dudoit S: **Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments.** *BMC bioinformatics* 2010, **11:**94.
17. Burrows M, Wheeler D: **A Block-sorting Lossless Data Compression Algorithm.** digital Systems Research Centre: 1994.
18. Campbell JL, Kleckner N: **E. coli oriC and the dnaA gene promoter are sequestered from dam methyltransferase following the passage of the chromosomal replication fork.** *Cell* 1990, **62:**967-979.
19. Chang WC, Li CW, Chen BS: **Quantitative inference of dynamic regulatory pathways via microarray data.** *BMC bioinformatics* 2005, **6:**44.
20. Cheung MS, Down TA, Latorre I, Ahringer J: **Systematic bias in high-throughput sequencing data and its correction by BEADS.** *Nucleic acids research* 2011.
21. Chinenov Y, Henzl M, Martin ME: **The alpha and beta subunits of the GA-binding protein form a stable heterodimer in solution. Revised model of heterotetrameric complex assembly.** *The Journal of biological chemistry* 2000, **275:**7749-7756.
22. Crick F: **Central dogma of molecular biology.** *Nature* 1970, **227:**561-563.
23. Das MK, Dai HK: **A survey of DNA motif finding algorithms.** *BMC bioinformatics* 2007, **8 Suppl 7:**S21.
24. **The GEM library** [http://sourceforge.net/apps/mediawiki/gemlibrary]
25. Dohm JC, Lottaz C, Borodina T, Himmelbauer H: **Substantial biases in ultra-short read data sets from high-throughput DNA sequencing.** *Nucleic acids research* 2008, **36:**e105.
26. Duftner N, Larkins-Ford J, Legendre M, Hofmann HA: **Efficacy of RNA amplification is dependent on sequence characteristics: implications for gene expression profiling using a cDNA microarray.** *Genomics* 2008, **91:**108-117.

27. Elnitski L, Jin VX, Farnham PJ, Jones SJ: **Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques.** *Genome research* 2006, **16:**1455-1464.

28. Elsner HI, Lindblad EB: **Ultrasonic degradation of DNA.** *DNA (Mary Ann Liebert, Inc* 1989, **8:**697-701.

29. Eshaghi M, Zhu L, Chu Z, Li J, Chan CS, Shahab A, Karuturi RK, Liu J: **Deconvolution of chromatin immunoprecipitation-microarray (ChIP-chip) analysis of MBF occupancies reveals the temporal recruitment of Rep2 at the MBF target genes.** *Eukaryotic cell* 2011, **10:**130-141.

30. Euskirchen GM, Rozowsky JS, Wei CL, Lee WH, Zhang ZD, Hartman S, Emanuelsson O, Stolc V, Weissman S, Gerstein MB, et al: **Mapping of transcription factor binding regions in mammalian cells by ChIP: comparison of array- and sequencing-based technologies.** *Genome research* 2007, **17:**898-909.

31. Fahlgren N, Howell MD, Kasschau KD, Chapman EJ, Sullivan CM, Cumbie JS, Givan SA, Law TF, Grant SR, Dangl JL, Carrington JC: **High-throughput sequencing of Arabidopsis microRNAs: evidence for frequent birth and death of MIRNA genes.** *PloS one* 2007, **2:**e219.

32. Feinberg AP, Vogelstein B: **Hypomethylation distinguishes genes of some human cancers from their normal counterparts.** *Nature* 1983, **301:**89-92.

33. Fisher RA: **Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population.** *Biometrika* 1915, **10:**507-521.

34. Flannery BP, Teukolsky SA, Vetterling WT: *Numerical Recipes in C: The Art of Scientific Computing.* Cambridge: William H. Press; 1988.

35. Freifelder D, Davison PF: **Studies on the sonic degradation of deoxyribonucleic acid.** *Biophysical journal* 1962, **2:**235-247.

36. Gentles AJ, Karlin S: **Genome-scale compositional comparisons in eukaryotes.** *Genome research* 2001, **11:**540-546.

37. Grokhovsky SL, Il'icheva IA, Nechipurenko DY, Golovkin MV, Panchenko LA, Polozov RV, Nechipurenko YD: **Sequence-specific ultrasonic cleavage of DNA.** *Biophysical journal* 2011, **100:**117-125.

38. Han JS, Kang S, Kim SH, Ko MJ, Hwang DS: **Binding of SeqA protein to hemi-methylated GATC sequences enhances their interaction and aggregation properties.** *The Journal of biological chemistry* 2004, **279:**30236-30243.

39. Han JS, Kang S, Lee H, Kim HK, Hwang DS: **Sequential binding of SeqA to paired hemi-methylated GATC sequences mediates formation of higher order complexes.** *The Journal of biological chemistry* 2003, **278:**34983-34989.
40. Hansen KD, Brenner SE, Dudoit S: **Biases in Illumina transcriptome sequencing caused by random hexamer priming.** *Nucleic acids research* 2010, **38:**e131.
41. Hooke R: *Micrographia or Some Physiological Descriptions of Minute Bodies* London: The Royal Society; 1665.
42. Horn PJ, Peterson CL: **Molecular biology. Chromatin higher order folding--wrapping up transcription.** *Science (New York, NY)* 2002, **297:**1824-1827.
43. Horowitz NH: **One-gene-one-enzyme: remembering biochemical genetics.** *Protein Sci* 1995, **4:**1017-1019.
44. Hower V, Evans SN, Pachter L: **Shape-based peak identification for ChIP-Seq.** *BMC bioinformatics* 2011, **12:**15.
45. Jensen ST, Liu XS, Zhou Q, Liu JS: **Discovery of Gene Regulatory Binding Motifs: A Bayesian Perspective.** *Statistical Science* 2004, **19:**188-204.
46. Ji H, Jiang H, Ma W, Johnson DS, Myers RM, Wong WH: **An integrated software system for analyzing ChIP-chip and ChIP-seq data.** *Nature biotechnology* 2008, **26:**1293-1300.
47. Jiang H, Wang F, Dyer NP, Wong WH: **CisGenome Browser: a flexible tool for genomic data visualization.** *Bioinformatics (Oxford, England)* 2010, **26:**1781-1782.
48. Johannsen W: *Elemente der exakten Erblichkeitslehre.* Jena: Fisher G.; 1909.
49. Johnson DS, Mortazavi A, Myers RM, Wold B: **Genome-wide mapping of in vivo protein-DNA interactions.** *Science (New York, NY)* 2007, **316:**1497-1502.
50. Kasowski M, Grubert F, Heffelfinger C, Hariharan M, Asabere A, Waszak SM, Habegger L, Rozowsky J, Shi M, Urban AE, et al: **Variation in transcription factor binding among humans.** *Science (New York, NY)* 2010, **328:**232-235.
51. Kel AE, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E: **MATCH: A tool for searching transcription factor binding sites in DNA sequences.** *Nucleic acids research* 2003, **31:**3576-3579.
52. Kharchenko PV, Tolstorukov MY, Park PJ: **Design and analysis of ChIP-seq experiments for DNA-binding proteins.** *Nature biotechnology* 2008, **26:**1351-1359.
53. Kircher M, Kelso J: **High-throughput DNA sequencing--concepts and limitations.** *Bioessays* 2010, **32:**524-536.

54. Klepper K, Sandve GK, Abul O, Johansen J, Drablos F: **Assessment of composite motif discovery methods.** *BMC bioinformatics* 2008, **9:**123.

55. Koohy H, Dyer NP, Reid JE, Koentges G, Ott S: **An alignment-free model for comparison of regulatory sequences.** *Bioinformatics (Oxford, England)* 2010, **26:**2391-2397.

56. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860-921.

57. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.** *Genome biology* 2009, **10:**R25.

58. Lawrence CE, Reilly AA: **An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences.** *Proteins* 1990, **7:**41-51.

59. Lee RC, Feinbaum RL, Ambros V: **The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14.** *Cell* 1993, **75:**843-854.

60. Li J, Jiang H, Wong WH: **Modeling non-uniformity in short-read rates in RNA-Seq data.** *Genome biology* 2010, **11:**R50.

61. Li R, Ackerman WEt, Summerfield TL, Yu L, Gulati P, Zhang J, Huang K, Romero R, Kniss DA: **Inflammatory gene regulatory networks in amnion cells following cytokine stimulation: translational systems approach to modeling human parturition.** *PloS one* 2011, **6:**e20560.

62. Li R, Li Y, Kristiansen K, Wang J: **SOAP: short oligonucleotide alignment program.** *Bioinformatics (Oxford, England)* 2008, **24:**713-714.

63. Lieb JD: **Genome-wide mapping of protein-DNA interactions by chromatin immunoprecipitation and DNA microarray hybridization.** *Methods in molecular biology (Clifton, NJ* 2003, **224:**99-109.

64. Liu JS, Neuwald AF, Lawrence CE: **Bayesian Models for Multiple Local Sequence Alignment and Gibbs Sampling Strategies.** *Journal of the American Statistical Association* 1995, **90:**1156-1170.

65. Locke G, Tolkunov D, Moqtaderi Z, Struhl K, Morozov AV: **High-throughput sequencing reveals a simple model of nucleosome energetics.** *Proceedings of the National Academy of Sciences of the United States of America* 2010, **107:**20998-21003.

66. Lozano JL, Sanchez JM, Mesa F: **A G2 framework for supervisory control of Ecosimpro experiments.** In *Electrotechnical Conference, 2006 MELECON 2006 IEEE Mediterranean*; *16-19 May 2006*. 2006: 409-412.

67. Lu M, Campbell JL, Boye E, Kleckner N: **SeqA: a negative modulator of replication initiation in E. coli.** *Cell* 1994, **77:**413-426.

68. Mamanova L, Andrews RM, James KD, Sheridan EM, Ellis PD, Langford CF, Ost TW, Collins JE, Turner DJ: **FRT-seq: amplification-free, strand-specific transcriptome sequencing.** *Nature methods* 2010, **7:**130-132.

69. Mann TL, Krull UJ: **The application of ultrasound as a rapid method to provide DNA fragments suitable for detection by DNA biosensors.** *Biosensors & bioelectronics* 2004, **20:**945-955.

70. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, et al: **TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes.** *Nucleic acids research* 2006, **34:**D108-110.

71. Mead EL, Sutherland RG, Verrall RE: **The effect of ultrasound on water in the presence of dissolved gases.** *Canadian Journal of Chemistry* 1976, **54:**1114-1120.

72. Moralee D: **The System X project** *Electronics and Power* 1979, **25:**544-551.

73. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq.** *Nature methods* 2008, **5:**621-628.

74. Nelder JA, Mead R: **A Simplex Method for Function Minimization.** *The Computer Journal* 1965, **7:**308-313.

75. Nilsen TW, Graveley BR: **Expansion of the eukaryotic proteome by alternative splicing.** *Nature* 2010, **463:**457-463.

76. Orlando V: **Mapping chromosomal proteins in vivo by formaldehyde-crosslinked-chromatin immunoprecipitation.** *Trends in biochemical sciences* 2000, **25:**99-104.

77. Park PJ: **ChIP-seq: advantages and challenges of a maturing technology.** *Nature reviews* 2009, **10:**669-680.

78. Pauli F, Myers R: **Myers Lab ChIP-seq Protocol, v041610.1 and v041610.2.** 2010.

79. Petraccone L, Erra E, Esposito V, Randazzo A, Mayol L, Nasti L, Barone G, Giancola C: **Stability and structure of telomeric DNA sequences forming quadruplexes containing four G-tetrads with different topological arrangements.** *Biochemistry* 2004, **43:**4877-4884.

80. Pillai RS: **MicroRNA function: multiple mechanisms for a tiny RNA?** *RNA (New York, NY* 2005, **11:**1753-1761.

81. Press WH, Teukolsky SA, Vetterling WT, Flannery BP: *Numerical Recipies in Fortran: The Art of Scientific Computing.* Cambridge: Cambridge University Press; 1992.

82. Reid J, Evans K, Dyer N, Wernisch L, Ott S: **Variable structure motifs for transcription factor binding sites.** *BMC genomics* 2010, **11:**30.

83. Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A, et al: **Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing.** *Nature methods* 2007, **4:**651-657.

84. Rosenbloom KR, Dreszer TR, Pheasant M, Barber GP, Meyer LR, Pohl A, Raney BJ, Wang T, Hinrichs AS, Zweig AS, et al: **ENCODE whole-genome data in the UCSC Genome Browser.** *Nucleic acids research* 2010, **38:**D620-D625.

85. Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, Bjornson R, Carriero N, Snyder M, Gerstein MB: **PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls.** *Nature biotechnology* 2009, **27:**66-75.

86. Sánchez-Romero MA, Busby SJW, Dyer NP, Millard AD, Grainger DC: **Dynamic Distribution of SeqA Protein across the Chromosome of Escherichia coli K-12.** *mBio* 2010, **1:**e00012-00010.

87. Schneider TD, Stephens RM: **Sequence logos: a new way to display consensus sequences.** *Nucleic acids research* 1990, **18:**6097-6100.

88. Schwartz S, Oren R, Ast G: **Detection and removal of biases in the analysis of next-generation sequencing reads.** *PloS one* 2011, **6:**e16685.

89. Schwarz G: **Estimating the Dimension of a Model.** *The Annals of Statistics* 1978, **6:**461-464.

90. Slominska M, Konopa G, Ostrowska J, Kedzierska B, Wegrzyn G, Wegrzyn A: **SeqA-mediated stimulation of a promoter activity by facilitating functions of a transcription activator.** *Molecular microbiology* 2003, **47:**1669-1679.

91. Spyrou C, Stark R, Lynch AG, Tavare S: **BayesPeak: Bayesian analysis of ChIP-seq data.** *BMC bioinformatics* 2009, **10:**299.

92. Stepanov A, Lee M: **The Standard Template Library.** Hewlett-Packard Company: 1995.

93. Suslick KS: **Sonochemistry.** *Science (New York, NY)* 1990, **247:**1439-1445.

94. Szalkowski AM, Schmid CD: **Rapid innovation in ChIP-seq peak-calling algorithms is outdistancing benchmarking efforts.** *Briefings in bioinformatics* 2010.

95. Teytelman L, Ozaydin B, Zill O, Lefrancois P, Snyder M, Rine J, Eisen MB: **Impact of chromatin structures on DNA processing for genomic analyses.** *PloS one* 2009, **4:**e6700.

96. Valouev A, Johnson DS, Sundquist A, Medina C, Anton E, Batzoglou S, Myers RM, Sidow A: **Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data.** *Nature methods* 2008, **5:**829-834.

97. Waldminghaus T, Skarstad K: **The Escherichia coli SeqA protein.** *Plasmid* 2009, **61:**141-150.

98. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics.** *Nature reviews* 2009, **10:**57-63.

99. Wang Z, Zang C, Cui K, Schones DE, Barski A, Peng W, Zhao K: **Genome-wide mapping of HATs and HDACs reveals distinct functions in active and inactive genes.** *Cell* 2009, **138:**1019-1031.

100. Watson JD, Crick FH: **Genetical implications of the structure of deoxyribonucleic acid.** *Nature* 1953, **171:**964-967.

101. Watson JD, Crick FH: **Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid.** *Nature* 1953, **171:**737-738.

102. Wolffe AP, Guschin D: **Review: chromatin structural features and targets that regulate transcription.** *Journal of structural biology* 2000, **129:**102-122.

103. Zhu L, Chou S-H, Reid BR: **The Structure of a Novel DNA Duplex Formed by Human Centromere d(TGGAA) Repeats with Possible Implications for Chromosome Attachment during Mitosis.** *Journal of molecular biology* 1995, **254:**623-637.

104. Zhu L, Zhang Y, Zhang W, Yang S, Chen JQ, Tian D: **Patterns of exon-intron architecture variation of genes in eukaryotic genomes.** *BMC genomics* 2009, **10:**47.